ABSTRACT

YUFANG BAO. Nonlinear Image Denoising Methodologies.
(Under the direction of Prof. Hamid Krim.)

In this thesis, we propose a theoretical as well as practical framework to combine geometric prior information to a statistical/probabilitstic methodology in the investigation of a denoising problem in its generic form together with its various applications in signal/image analysis.

We are able in the process, to investigate, understand and mitigate existing limitations of so-called nonlinear diffusion techniques (such as the Perona-Malik equation) from a probabilistic view point, and propose a new nonlinear denoising method that is based on a random walk whose transition probabilities are selected by the information of a two-sided gradient. This results in a piecewise constant filtered image and lifts the long-standing problem of an unknown evolution stopping time.

Our second contribution is in establishing a direct link between multi-resolution analysis techniques and so-called scale space analysis methods, which we in turn utilize to improve the performance of segmentation-optimized image analysis techniques. This is accomplished by using wavelets of higher order vanishing moments, specifically, we achieve a reduction in the typical "blocky" artifacts and a better preservation of texture information.

Our third and final contribution is to propose a drastically different approach by isolating statistically independent components in a signal, which we later use as a basis for discrimination against noise, or potentially as plain features. This is related to the well known independent component analysis (ICA), for which we first propose $\alpha-$Jensen -Rényi divergence as an information- theoretic criterion. In addition, we propose a Rényi mutual divergence as a better criterion to separate mixed signals along with a non-parametric estimation technique for such a measure for 1-D problems.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2002		**00-05-2002 to 00-05-2002**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Nonlinear Image Denoising Methodologies	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **North Carolina State University,Department of Electrical and Computer Engineering,Raleigh,NC,27695**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**130**	

NONLINEAR IMAGE DENOISING METHODOLOGIES

By

Yufang Bao

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND COMPUTER ENGINEERING
AT
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NORTH CAROLINA
MAY 2002

Approved by

Prof. Hamid Krim Chair of Advisory Committee	Prof. Arne A. Nilsson ECE Dept. NCSU
Prof. Alexandra Duel-Hallen ECE Dept. NCSU	Prof. Zhilin Li MATH. DEPT NCSU
Dr. Robert Cohen SAS Institute	

To My husband Zhenwu

and

My sons Suzhou and David Suyuan

Biography

Yufang Bao received her first Ph.D. degree in mathematics from Beijing Normal University, Beijing, China in 1995. She also received her B.S. and M.S. degrees in mathematics from Fujian Normal University, Fuzhou, China, in 1989 and 1992 respectively. From 1995 to 1998, she worked as a university instructor in Beijing University of Posts and Telecommunications. In April 1998, she started visiting North Carolina State University, Raleigh, NC and later became a student for her second Ph.D. degree in January 1999. Her research interests are in statistical signal/image processing and stochastic modelling.

Acknowledgements

First and foremost, I would like to thank Prof. Hamid Krim. for his tireless support throughout my Ph.D. education. I greatly benefited from enthusiastic discussions with him and learned from every piece of the abundant literature references he recommended. In every sense, none of this work would have been possible without him. He was a constant source of inspiration, including the visit to MIT he had arranged for me, and which I greatly learned from.

Many thanks go to Prof. Arne. A. Nilsson, Prof. Alexandra Duel-Hallen and Prof. William McEneaney for their recommendations in the course of my job search. I would especially like to express my gratitude to Prof. Nilsson for his encouragement and support through some difficult time. I also want to thank Prof. Duel-Hallen for her kind support, and her encouraging smile. I need to thank Prof. McEneaney, who was on my committee but no longer available due to his move, I benefited from a number of discussions with him. I would also like to thank Prof. Zhilin Li, who kindly joined my committee to replace Prof. McEneaney, and whom I had the pleasure to first meet when he gave an inspiring seminar to our group VISSTA. I am thankful to Dr. Robert Cohen, for his comments as well as his sincere support. I would also need to extend my thanks to Prof. Wesley Snyder and Prof. Joel Trussell for their help.

Far too many people to mention individually have assisted me during my research work at NCSU, I would like to thank them here, in particular, I would like to thank my colleagues for their help.

I would like to thank Ms. Sandy Bronson, for her kind help.

The financial supports from both Air Force Office of Scientific Research (AFOSR) and the ECE Dept. of NCSU are greatly appreciated.

I am grateful to my family for their encouragement. Special thanks go to my husband, Zhenwu Li, for his endless love, support and encouragement, and also to my son Suzhou's understanding and help, and to our baby, David Suyuan Li, he makes life sunny for us.

Contents

List of Figures **ix**

1 Introduction **1**

1.1 Problem Motivation and Formulation 1

1.2 Summary of Thesis Main Contributions and Organization 4

2 Scale Space, Diffusion and Variation **6**

2.1 Scale Space Concept 6

2.2 Probabilistic View of Diffusion 8

2.2.1 Diffusion and SDE 8

2.2.2 Kolmogorov's Backward and Forward Equations 9

2.2.3 Example: Brownian Motion 12

2.3 Variational Methodology 13

2.4 Appendix 15

3 Nonlinear controlled diffusion **17**

3.1 Definition of Nonlinear Controlled Diffusion 17

3.2 Nonlinear Diffusion and PDE 18

3.3 Nonlinear Diffusion on a Lattice 20

3.3.1 Discrete Approximation of Diffusion 20

3.3.2 Finite Markov Chain Example 22

3.3.3 Discrete Time/Scale Evolution 23

3.3.4 A Stochastic View of Perona-Malik Equation 24

3.4 Two Sided Gradient-Driven Diffusion 27

3.5 Discussion 29
3.6 Experimental Results 30
3.7 Conclusion 32
3.8 Appendix A 33

4 Multiscale Wavelet space **40**
4.1 Definition of a Wavelet 40
4.2 Wavelet frames 42
4.3 Wavelet basis 43
4.3.1 Wavelet design: Connection to conjugate mirror filters 45
4.3.2 Vanishing Moments vs. Support size of a wavelet 47
4.4 Daubechies Wavelets 48
4.5 Wavelet Packet 49

5 Wavelet Frame-Based Nonlinear Filtering **52**
5.1 Introduction 52
5.2 Problem Statement 53
5.3 A Multiscale Approach to Scale-Space Analysis 53
5.4 Frame Representation and Reconstruction 54
5.5 Selection and Impact of a Wavelet Support 58
5.6 Image Reconstruction using a Haar Frame 63
5.7 Smoothing in the Frame Domain 65
5.8 Nonlinear Reconstruction 66
5.9 Experimental Results 68
5.10 Appendix 73

6 Independent Component Analysis **75**
6.1 Linear ICA models 75
6.2 Existing ICA Algorithms 78
6.3 Applications of ICA 82

7 New measure criteria for ICA **83**

7.1 Definition of Rényi Entropy 83

7.2 Jensen-Rényi divergence as a new criterion for ICA 84

7.2.1 Introduction to $\alpha-$Jensen-Rényi divergence 84

7.2.2 $\alpha-$Jensen-Rényi divergence as an Independence Measure 86

7.2.3 Application to ICA 86

7.3 $\alpha-$Rényi mutual divergence as a New Criterion for ICA 88

7.3.1 Introduction to $\alpha-$Rényi divergence 88

7.3.2 Comparison between $\alpha-$Rényi mutual divergence and Mutual Information 92

7.4 Application to ICA 94

7.5 Approximation of $\alpha-$Rényi Mutual Divergence 96

7.5.1 Introduction 96

7.5.2 Approximation Theorems of $\alpha-$Rényi mutual divergence 96

7.6 Conclusion 101

7.7 Appendix A. 102

7.8 Appendix B 106

8 Possible Future Work **109**

8.1 Summaries 109

8.2 Possible Future Research 110

Bibliography **111**

List of Figures

2.1 Sample pathes of a random process that begin at x at time s and arrive at y at time t . . . 11
2.2 Random process sample pathes that begin with position y at time t and arrive in x at time s at the inverse time direction . . . 12
2.3 A particle (pixel) may diffuse over many possible paths, and an average is usually computed. . . . 13

3.1 Finite Controlled Markov Chain modelling. . . . 23
3.2 Noisy signal filtered by our random walk algorithm and PM algorithm. . . . 31
3.3 Stable signal remains unchanged following proposed nonlinear diffusion. . . . 32
3.4 A noisy image together with its enhanced copy by the proposed algorithm and by the P-M method best result. . . . 34
3.5 Complete Smoothing vs Stability. . . . 35
3.6 PM algorithm. . . . 36
3.7 Checker Board Enhancement. . . . 36
3.8 tools segmentation. . . . 37
3.9 House segmentation. . . . 37
3.10 Circle segmentation. . . . 38
3.11 Error rate in circle segmentation for different SNR scenarios. . . . 39

4.1 Daubechies scaling function ϕ (top) and wavelet ψ(bottom) with vanishing moments 2 . . . 50

5.1 A profile of noisy Lenna image and filtered result with Random walk algorithm[52]. 54

5.2 Spectral characteristics of Heat equation-like filter using Haar and Daubechies-4/6 wavelet functions. 59
5.3 A discontinuous signal and its decomposition as sum of continuous part and discontinuous part, also approximation of the continuous part. 62
5.4 A profile take from the rock texture image, with filtered result 63
5.5 One possible nonlinear functional is an exponential weighting. 67
5.6 A noisy Lenna image and filtered result with three algorithms. 69
5.7 A texture image, noisy texture image and filtered result with Daubechies 4 . 70
5.8 A texture image, noisy fabric image and filtered result with Daubechies 4 . . 71
5.9 A texture with rocks image, its noisy image and filtered result with Daubechies 4 72

6.1 Independent source signals and mixed signals obtained by rotation. 77

7.1 Renyi entropy of Bernoulli distributions at several α compared to Shannon entropy 85
7.2 An application using α-JR Divergence 89
7.3 ICA criterion using mutual information and α-JR divergence. 89
7.4 An application using α-JR Divergence 90
7.5 ICA criterion using mutual information and α-JR divergence. 91
7.6 Approximated 0.2−Rényi mutual divergence and its exact theoretical value compared to mutual information 94
7.7 Approximated 1.8−Rényi mutual divergence and its exact theoretical value compared to mutual information 95
7.8 Mixed signals and its separation using 1.6−Rényi mutual divergence. 97
7.9 1.6−Rényi mutual divergence measure compared to mutual information . . . 98
7.10 Example: Two nested partitions of $\mathcal{R}^2$ 100
7.11 Approximated 0.5−Rényi mutual divergence and its exact theoretical value 102
7.12 Functions of $f = x^\alpha$, with $0 < \alpha < 1$ and $1 < \alpha \leq 2$ 103

Chapter 1

Introduction

One/two dimensional signals characterized by singularities are usually contaminated by additive noise, which cause difficulties in localizing them. This thesis will address this issue as related to problems of signal restoration, segmentation and edge detection as briefly described in this chapter. Upon motivating and formulating the basic problem, we summarize our contributions in this direction using a stochastic random walk, wavelet frame theory and information measure as the basic analytical tools.

1.1 Problem Motivation and Formulation

The primary goal of processing a noisy signal is to obtain a reconstruction as close to the original clean signal as possible, which, in turn, provides a reliable(hopefully, robust) version for segmentation, and edge detection of a signal/image. The design of a filter targeted for denoising purpose is normally based on some prior knowledge about the signal, e.g., staircase or smooth etc. In this thesis, Our approach to denoising is first based on a controlled nonlinear stochastic random walk to achieve a scale space analysis(as in Chapter 2, 3) to enhance images. To better preserve texture in images, in Chapter 5, we use wavelet frames to simultaneously improve the enhancement as well as the segmentation. In chapter 7, we introduce two new information theoretical approaches to extract independent components from mixed signals.

A classical method to restore a useful signal is to adapt a probabilistic signal prior model in applying a statistical methodology, such as Maximum A Posteriori(MAP) estimation, see[32, 31]. A precise probabilistic model is, however, usually unavailable and a wrong

model may yield significant errors and is unacceptable in signal recovery problem. An effort of adopting a more objective energy functional maybe constructed to derive a MAP-like principle by line of variational formulation approach and rooted in the intrinsic geometry and smoothness of the signal. The optimization of such a functional by way of the Euler-Lagrange equation yields a steepest gradient descent search for the optimal signal/image. The so-obtained partial differential equation (PDE) is an evolution of a signal/image through scales.

Scale space, first introduced as a homogenous dynamic low-pass filter linearly smooth away noise with increasing of scale, was extended to a nonlinear selective smoothing by Perona-Malik [74, 75, 89]. This triggered an intense interest in searching for new nonlinear filters to better preserve features [73, 78, 93, 87, 60, 94, 15, 58]. While simple to implement, these procedures become complex involved for noise contaminated images[8, 91]. Several improvements have been proposed since, and for example, a more flexible technique, which has been shown to be equivalent to anisotropic equation is that of Mean-field annealing(MFA)[40], for which, the parameter choice is much simpler.

Most, if not all, of existing techniques have been predominantly deterministic in nature, with little or no stochastic treatment or interpretation of the diffusion. In addition, unless a specific stopping time is known to be adequate, the resulting evolution equation is well known to almost always lead to a complete smoothing of the signals(i.e., the steady state of the PDE). Pollak *et. al.* [77, 78] recently proposed an approach addressing robustness issues, and showed some remarkable results for a wide class of perturbation noises. The analysis remained as in all other cases, fundamentally deterministic, and also required knowledge of the stopping time for the evolution.

From a probabilistic vantage point, the characteristics of the random process which underly the diffusion have so far been overlooked, and their overall influence on the solution in different scenarios has remained unclear. One of our goals in this thesis is to first detail a probabilistic framework which helps us provide an alternative view of the nonlinear diffusion problem. This in turn, is instrumental in our providing an alternative interpretation of existing methods, e.g. Perona-Malik equation, and in using the gained insight to propose a solution to its well known limitations. More specifically, we view an evolution equation by way of a controlled diffusion [57] strategies as a solution resulting from an optimization of an

energy functional. This ultimately leads to a two/four state Markov Chain (MC) with one step transition probabilities well adapted to preserving the salient features of a signal/image(such as edges) while smoothing away the noise. As will be elaborated on further below, in addition to a marked performance improvement over P-M equation, and by way of our newly proposed technique, we are able to lift a longstanding problem in nonlinear diffusion, namely requiring to have prior knowledge of the stopping time. We in fact show that the stable point for our equation is a staircase function.

The resulting image as a staircase function is, however, at a cost of a loss of texture in the image. This as further elaborated on below, is inherent to the first order Markov property assumed for the image and implicit in the edge modelling(by a first order difference gradient). In the second part of our work, we address the texture loss problem in the course of smoothing by the before mentioned technique.

As we can see, piecewise picture is the best effort we may obtain through evolution so far. On the other hand, wavelet theory provides various methods to explore the intrinsic properties of a signal, wavelets of higher order vanishing moments result in fewer large detail coefficients if a function is smooth, and the decomposition of a signal into a wavelet frame reveals redundant information. We also can see that wavelet packets provide a tool to explore more detail content from spectral domain of view. Inspired by these facts, a nature question then arose is whether we can investigate the interplay between PDE-based filtering and multiscale analysis. This promising idea is implemented and a texture preserved algorithm is proposed as shown in Chapter 5.

We show that using frames of wavelet of higher order vanishing moments than Haar's is tantamount to accounting for longer term correlation structure, while preserving the local focus on equally important features(e.g. edges). This hence yields an efficient tool in analyzing and in enhancing images with a careful account for texture information. We propose to decompose images into Daubechies 4 based wavelet frames, where redundant information will be generated. That information is useful when we deal with noisy image although it is not necessary when we try to reconstruct image. The problem is that, we would like to recover a clear, enhanced version of the original one if the picture given is noisy. We maintain that potentially useful information lies in the redundant representation of a signal/image and should be fully exploited.

Our third approach to separating and localizing various components of a signal process is to ensure a statistical independence among them, thereby also affording one to extract features of importance. This will be achieved by seeking to extract independent components and whose higher order components is better regarded, as it better separate signals from others. While many approaches spanning higher order statistics to learning algorithms have been proposed, we proposed two new information measures which depend on the probability density functions.

With a improve of a non-parametric estimation of mutual information, we propose a non-parametric Rényi mutual divergence approximation using dependent data, which, together with the better measurement property of Rényi mutual divergence over mutual information enable us to practically apply it as an alternative criterion to ICA. We also propose using $\alpha-$Jensen-Rényi divergence that was recently developed ([37, 1]) as a new independent measure among more than two pdf's in lieu of mutual information. We show it to improve performance in separating sources[5], as one way impose weighting priors (hence contribution) of different data sets.

1.2 Summary of Thesis Main Contributions and Organization

In Chapter 3 we propose a stochastic framework where nonlinear diffusions are cast and are given an insightful interpretation towards understanding their intrinsic behaviors. We reinterpret a linear evolution partial differential equation(PDE) as a direct result of a mean value of random walk functional. In addition, Perona-Malik equation is also interpreted as a controlled random walk for which an adjusted energy functional yields a much improved algorithm. This in fact results in a new diffusion method based on two-sided gradient is proposed in section 2.6, which yields piecewise constant filtered images. Additional extensions were also proposed [49].

In Chapter 5, We propose a new evolution-equation based technique that utilizes multi-resolution wavelet frame coefficients. Wavelet frame coefficients include redundant information and when the wavelets are of higher order vanishing moments, a longer correlation structure is account for. We provide a brief contextualization and statement of the problem. An explanation of the decomposition and reconstruction is given in section 4.3, where we

also prove that the Heat diffusion is equivalent to subtracting second level detail of Haar frame coefficients. In section 4.4, we explain from a spectral perspective the effect of a vanishing moment of a wavelet and proceed to derive the detailed implementation equations. We finally provide some substantiating denoising image examples as a conclusion in 4.8. The major results of this chapter have been published in [4, 6].

In Chapter 6, we provide a brief review independent component analysis (ICA) and discuss various contrast functions used to recover independent source signals. It is shown that all the contrast functions are in fact related to mutual information and the MLE principle. This subsequently leads us to propose new information criteria, such as JR divergence and Rényi mutual divergence –Examples are also provided.

In Chapter 7, We expound on these measures, show their application to ICA and develop non-parametric technique for approximating the Rényi mutual divergence by a cell approximation algorithm.

In Chapter 8, we provide some extensions and new research areas for future work.

Chapter 2

Scale Space, Diffusion and Variation

Scale space, known as a collection of signals output from a dynamic filter whose transform functions varied with time/scale, provides a flexible choice to meet different processing purposes by specifying a time/scale. Scale space filtering is effected via a partial differential equation(PDE), which, as we will see in the following, is related to an underlying particle's motion which is in turn governed by a stochastic differential equation(SDE). Our motivation of the scale space methodology proceedly exploiting the tight connection between a PDE and a SDE, and the subsequent stochastic interpretation of the PDE solution. The interplay between a PDE and an energy functional yields a deep insight, which in turn as we will elaborate in later chapters, leads to a clarifying and solving some outstanding problems.

2.1 Scale Space Concept

Scale-based analysis has recently played an increasingly important role in signal and image analysis since Witkin's ground breaking paper[95], in which a so-called linear scale space was constructed and the following linear evolution partial differential equation(PDE) effecting the filtering was proposed

$$\begin{aligned} \frac{\partial U(t,x)}{\partial t} &= \Delta U(t,x) \\ U(0,x) &= f(x). \end{aligned} \tag{2.1.1}$$

The symbol Δ denotes a Laplacian operator acting on filtered signals $U(t,x)$, which may be interpreted as copies of an original signal $U(0,x)$ at a fixed time/scale t. This was based on the conclusion that convolving a signal with a Gaussian kernel was equivalent to evolving it

with a Heat differential operator as shown in the next equation, where time now plays the role of scale [48, 97],

$$U(t,x) = \frac{1}{(\sqrt{(2\pi t)})^d} \exp\left(-\frac{|x|^2}{2t}\right) * U(0,x)$$

where $x = (x_1, \cdots, x_d) \in \mathcal{R}^d,\ |x| = \sqrt{\sum_{i=1}^{d} x_i^2}$, $'*'$ denotes convolution of two functions. From a frequency domain's viewpoint, this equation is equivalent to

$$\hat{U}(t,\omega) = \hat{U}(0,\omega) exp(-t|\omega|^2).$$

where $\hat{U}(t,\omega)$ represents the spatial Fourier Transform of $U(t,x)$. This clearly explains that the high frequency content, which represents sharp features as well as noise, is removed when t increases or the scale becomes coarser and coarser, i.e., both noise and details will be smoothed out and no new information added.

The interest in scale space analysis stems from the fact that optimally processing image features may be tracked across the scale. The latter may be made to vary nonlinearly with a proper modification of the Gaussian kernel.

Linear heat diffusion, first introduced as a homogenous scale space in filtering theory by Witkin([95]), is an isotropic method that smooth signals with Gaussian kernel uniformly by increasing scale, which removes important features along with noise, this lead to the desire of scale space approaches that naturally preserve the intra-scale correlation information. An approach deployed such scale information was first proposed by Perona and Malik in their landmark paper([74]) and was aimed at preserving important sharp features such as edges. Their technique may also be viewed as a nonlinear filter whose selective smoothing is based upon the computed local gradient (maximal smoothing in low gradient or homogeneous regions, and minimal smoothing in high gradient regions), where signals are viewed as nothing but piecewise constant functions. The novelty of this approach together with its very promising results triggered a tremendous research activity in computer vision and applied mathematics [73, 78, 93, 87, 60, 94], where its mathematical properties as well as its numerical implementations and applications were investigated. A slight regularization by a Gaussian kernel to finest smooth a noisy signal was proposed in [15, 58] prior to implementing

nonlinear selective smoothing on the signal. On the other hand, this was interpreted in [63] as a robust estimation that resulted in an edge-stopping function to be applied to gradient. Many approaches have been proposed to address a variety of issues specific to images, such as, image enhancement, segmentation, and edge detection, which have been figured among the most often studied on account of their great relevance to low-level vision.[74, 75] (see [89] for a good review of the literature).

2.2 Probabilistic View of Diffusion

The fact that scale space analysis is defined by a PDE-based diffusion is essential to its probabilistic interpretation. Diffusion is used to describe a physical phenomenon that governs the transport of heat flow moving from a high to a low spatial concentration, This may in turn be investigated as a stochastic process of an underlying particle's movement. This is described by a stochastic differential equation(SDE) which has a corresponding macroscopic (by the theory of large numbers) manifestation by way of a PDE as discussed next.

2.2.1 Diffusion and SDE

A stochastic process $X_t(x)$ may be defined as a parameterized collection of random variables $\{X_t(x)\}_{t \in [0,T]}$ defined on a probability space $(\Omega,\ \mathcal{F},\ P)$ and assuming values in $\mathcal{R}^n$ in general, so that,

$$\forall\ \omega\ \in\ \Omega,\ \omega\ \rightarrow\ X_t(x,\omega),\ t \in\ [0,T],$$

where Ω is the usual sample space, $\mathcal{F}$ the σ-field and P the probability measure. A nice and intuitively appealing interpretation for ω is that of a moving particle whose starting position is x at time 0 and whose position at time t is given by $X_t(x,\omega)$. We will write X_{st} or $X_{st}(x)$ to denote a random process starting at x at time s, and currently at time t.

Definition 1. *Let $b(t,x)$ and $\sigma(t,x)$ be continuous in t,x and assume that for some constant $K \in R$*

$$|b(t,x)|^2 + |\sigma(t,x)|^2 \leq K(1+|x|^2) \tag{2.2.1}$$

and that for each $N \in R$, $\exists\ L_N$ with $|x| \leq N, |y| \leq N$, for which

$$|b(t,x) - b(t,y)| + |\sigma(t,x) - \sigma(t,y)| \leq L_N|x-y|. \tag{2.2.2}$$

A d-dimensional stochastic process $X_t = (X_t^1, \cdots, X_t^d)^T$ that satisfies the following stochastic differential equation(SDE) exists and is called an Ito-diffusion[71].

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t, \tag{2.2.3}$$

where B_t is a dimension m standard Brownian motion vector and $b(t, x)$ is a $d \times 1$ drift coefficient vector and $\sigma(t, x)$ is a $d \times m$ diffusion coefficient matrix(in this paper, we normally consider $d = 1$, $d = 2$ and $m = 1$ case).

Diffusion processes defined above have continuous paths, and when $\sigma(t, x) = \sigma(x)$, $b(t, x) = b(x)$, the diffusion processes are homogenous processes. We denote by $p(s, x, t, dy)$ the probability transition function of a stochastic process X_t, i.e., $p(s, x, t, dy) = P(X_t \in dy | X_s = x)$, and by $p(s, x, t, y)$ the probability transition density function of a particle starting at location x at time s and reaching y at time t. As described by the following theorem, the transition probability is normally determined by the drift and the diffusion coefficients, which characterize how the diffusion behaves as well. The infinitesimal generator (i.e., continuous operator which describes such a motion) of the diffusion in Eq. (2.2.3) can then be written as :

$$L_t = \frac{1}{2}\sum_{i,j=1}^{d} a_{ij}(t, x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(t, x)\frac{\partial}{\partial x_i} \tag{2.2.4}$$

where $a(t, x) = \sigma(t, x)\sigma(t, x)^T$.

2.2.2 Kolmogorov's Backward and Forward Equations

While our focus herein is on clarifying the situations which needed to be consider for unravelling the connection between a diffusion process and its corresponding PDE, the details of the theorems below may be found in [3, 29, 34].

Theorem 1. *Let $p(s, x, t, y)$ be the transition probability density function of diffusion process $X_t, 0 \le t \le T$ with continuous coefficients $b(t, x), \sigma(t, x)$ that satisfy certain conditions, then p is a so-called fundamental solution of the Kolmogorov's backward equation with L_s given in Eq. (2.2.4),*

$$\begin{aligned} &\frac{\partial p}{\partial s} + L_s p = 0 \\ &\lim_{s \uparrow t} p(s, x, t, y) \;\; = \;\; \delta(x - y) \end{aligned} \tag{2.2.5}$$

Theorem 2. *Let a transition density function $p(s, x, t, y)$ of a diffusion process satisfy certain conditions, and*

$$\frac{\partial p}{\partial t}, \quad \frac{\partial (b_i(t, y)p)}{\partial y_i}, \quad \frac{\partial^2 (\sigma(t, y)p)}{\partial y_i \partial y_j} \tag{2.2.6}$$

exist and be continuous, then $p(s,x,t,y)$ is a fundamental solution of Kolmogorov's forward equation for fixed s and x such that $s \leq t$.

$$\begin{aligned} \frac{\partial p}{\partial t} &= L_t^* p \\ \lim_{t \downarrow s} p(s,x,t,y) &= \delta(x-y) \end{aligned} \tag{2.2.7}$$

where L^* is the adjoint operator of L given by

$$L_t^* p = \sum_{i,j=1}^{d} \frac{\partial^2 (a_{ij}(t,y)p)}{\partial y_i \partial y_j} - \sum_{i=1}^{d} \frac{\partial (b_i(t,y)p)}{\partial y_i} \tag{2.2.8}$$

According to these two fundamental solutions of Kolmogorov's equations, the following probabilistic solutions of Kolmogorov's equations(PDEs) are formulated in term of the initial conditions.

Theorem 3. *[25] Assume f and L_t satisfy certain technical conditions, then $U(t,x) = E\{f(X_{s,t}(x))\} = \int f(y)P(s,x,t,y)dy$, $s < t$, is the solution of the following PDE that has an infinitesimal operator L_t as in Eq. (2.2.4),*

$$\begin{aligned} \frac{\partial U(t,x)}{\partial t} &= L_t U(t,x) \\ U(s,x) &= f(x). \end{aligned} \tag{2.2.9}$$

Theorem 4. *[34] Assume f and L_t satisfy certain technical conditions, then $U(s,x) = E\{f(X_{s,t}(x))\} = \int f(y)P(s,x,t,y)dy$, where $s < t$, is the solution of the following PDE that has an infinitesimal operator L_s as in Eq. (2.2.4),*

$$\begin{aligned} \frac{\partial U(s,x)}{\partial s} + L_s U(s,x) &= 0 \\ U(s,x) &= f(x) \quad \text{for } s \uparrow t. \end{aligned} \tag{2.2.10}$$

Theorem 5. *[34] Assume f and L_t satisfy certain technical conditions, then $U(t,y) = \int f(x)P(s,x,t,y)dx$, $s < t$, then $U(t,y)$ is the solution of the following PDE that has an infinitesimal operator L_t^* as in Eq..(2.2.8),*

$$\begin{aligned} \frac{\partial U(t,y)}{\partial t} &= L_t^* U(t,y) \\ U(t,y) \to f(y) &\qquad \text{for } t \downarrow s. \end{aligned} \tag{2.2.11}$$

With a closer look at the solution of Eq. (2.2.11), we can see that the solution can not be expressed as a mean value of a random process since it is integrated over all the initial

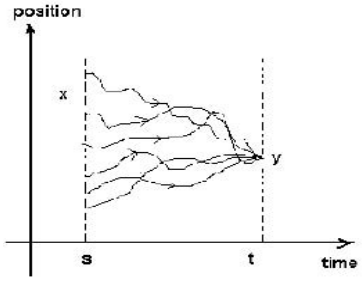

Figure 2.1: Sample pathes of a random process that begin at x at time s and arrive at y at time t

positions and is in contrast to the case in Th. 6. In order to clarify the difference, we provide two figures, one of which (fig. 2.1) corresponds to a solution of Eq. (2.2.11).

Next, we consider diffusions where the evolution time direction is reversed, a so-called backward diffusion(see fig.(2.2) and reference [55][54]), for which an alternative form of Kolmogorov's backward equation is given as

Theorem 6. *Assume that f and L_t satisfy certain technical conditions. The expected value $U(t,x) = E\{f(\hat{X}_{s,t}(x))\}$ is the solution to the following backward PDE on R^d*

$$\begin{aligned} \frac{\partial U(t,x)}{\partial t} + L'_t U(t,x) &= 0 \\ U(t,x) &= f(x) \quad \textit{for } t \downarrow s \end{aligned} \tag{2.2.12}$$

where L'_t is an alternative adjoint operator of L_t and is defined as

$$L'_t = \sum_{i,j=1}^{d} a_{ij}(t,x)\frac{\partial^2}{\partial x_i \partial x_j} - \sum_{i,j=1}^{d} b_i(t,x)\frac{\partial}{\partial x_i} \tag{2.2.13}$$

and $\hat{X}_{s,t}, s < t$ is a backward diffusion process, it also denoted as $\hat{X}_s(x)$ with the initial condition $\hat{X}_T = x$. it can also be defined as a forward diffusion per Definition 1 by writing $\hat{X}_s(x) = X_{T-s}(x) = X_t, 0 < t < T$, $\hat{X}_s(x)$ satisfies the following stochastic integral equation,

$$\hat{X}_{s,t}(x) = -\int_s^t b(r, \hat{X}_{r,t}(x))dr + \sum_{k=1}^{d}\int_s^t \sigma_k(r, \hat{X}_{r,t}(x))d\hat{B}_r^k$$

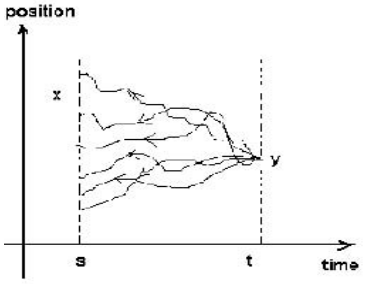

Figure 2.2: Random process sample pathes that begin with position y at time t and arrive in x at time s at the inverse time direction

2.2.3 Example: Brownian Motion

An illustrating example of a linear diffusion is the process described by the PDE in Eq. (2.1.1), in which L is specified as a Laplacian operator Δ (i.e., $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$). Brownian Motion, clearly a homogeneous random processes, therefore yields a transition probability density function denoted by $p(t, x, y) = p(s, x, s + t, y)$, where x and y may be exchanged as a result of the symmetry property of this diffusion. This property is rare for most of other processes. As noted earlier, diffusion of heat in a homogeneous medium fundamentally stems from the motion of particles, and it can be shown that the inherent randomness of this motion is well-described by a Brownian motion B_t [71], where an individual outcome $\omega \in \Omega$ in the prevailing sample space, may be associated to a particle. The process B_t may then be interpreted as, originating at time 0 from position x (assumed 0 for simplicity), the distance travelled by particle ω at time t. It is well known that a transition probability density for a Brownian motion in 1-D case, for instance, is a Gaussian PDF $p(t, x, y) = \frac{1}{(2\pi t)^{1/2}} e^{-\frac{(y-x)^2}{2t}} \quad \forall\ x,\ y \in \mathcal{R},\ t > 0$ (recall that a Brownian motion has independent Gaussian increments). It is thus clear that a stochastic interpretation of a solution (if it exists) subjected to some differentiability conditions, can be given by way of an ensemble average [25]

$$U(t, x) = E_x\{f(B_t)\} = \int_{\mathcal{R}} p(t, x, y) f(y) dy, \tag{2.2.14}$$

where the expectation $E(\cdot)$ is computed over all possible reachable positions y starting at position x. In the 2-D case, it is similarly possible to have such an interpretation as displayed

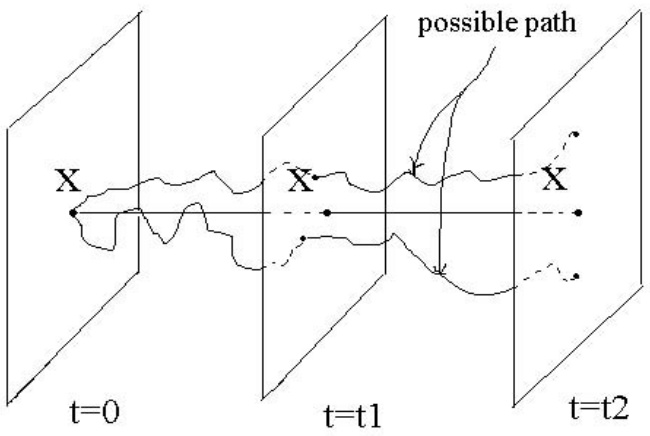

Figure 2.3: A particle (pixel) may diffuse over many possible paths, and an average is usually computed.

in Fig. 2.3. The times $t = t_1$ and $t = t_2$ are the instants at which all possible positions are averaged to yield a solution at the respective times.

2.3 Variational Methodology

While variational methods have been investigated in problems where minimizing cost is of interest, i.e., energy functionals derived from a MAP principle[32, 31, 35], it has been recently adopted to explain the mathematical foundations of scale space analysis from an optimization theoretic viewpoint[62]. Specifically, many existing evolution equations were shown to result from a minimization of energy functionals. The resulting Euler-Lagrange equations lead to a steepest gradient descent method giving rise to a PDE. Using this approach, we can establish the well known result that the Brownian motion is a result of minimizing

$$E(u) = \int |\nabla u|^2 dx, \tag{2.3.1}$$

The resulting PDE as given in Eq. (2.1.1) generates different copies of the image, at different scales, and in light of the above probabilistic interpretation, effects a particle motion

from a region of high density to one of low density. Additional insight is achieved by way of the following theorem applied to a generic functional.

Theorem 7. *For an energy function that takes the form*

$$E(u) = \int f(u, \nabla u) dx \tag{2.3.2}$$

where $f(u, \nabla u)$ *is given as*

$$f(u, \nabla u) = \frac{1}{2} \sum_{i,j=1}^{d} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} + \sum_{i=1}^{d} b_i u \frac{\partial u}{\partial x_i} \tag{2.3.3}$$

with $a_{ij} = a_{ji}$ *and* a_{ij} *satisfy certain conditions as in definition 1, then the resulting PDE from steepest gradient descent method is given as Eq. (2.2.9). This means that the underlying particle motion is homogenous and governed by a SDE described by an infinitesimal operator* L_t *of Eq. (2.2.4) with* $a_{ij}(t, x) = a_{ij}, b_i(t, x) = b_i$ *and* $a_{ij} = a_{ji}$.

Proof: See Appendix.

The infinitesimal operator as in Eq. (2.2.4) now further invokes a gradient of the function of interest. This is reflected by the energy functional and its precise effect is only clear when $a_{ij} > 0$, $i = j$ and $a_{ij} = 0$, $i \neq j$, (note that additional interactions among the component are present when $a_{ij} \neq 0$, $i \neq j$), and $b_i = 0$ (as in the Laplacian case).

One may, however, consider other generalizations (general $a(t, x)$, $b(t, x)$) for more elaborate effects as a function of scale and space. A ease in point is that of avoiding the trivial smoothing of an image resulting from a linear heat equation, and that of rather presenting key features through nonlinear transformations to slow down/eliminate some specific filtering.

A number of very good papers have provided inspiring variational interpretations to various nonlinear smoothing techniques [89, 83, 96, 60] and thus proposed their specific generalized denoising methods. In [83] such an approach resulted in a constrained total variation with a gradient descent formula. A very clear explanation of nonlinear diffusion resulting from variational methodology is given in [96].

2.4 Appendix

Proof of Theorem 7:

Note that $f_i'(y_0, y_1, \cdots, y_d) = \dfrac{\partial f}{\partial y_i}$, $i = 0, 1, \cdots, d$. For $u(x)$, $x = (x_1, x_2, \cdots, x_d)$, $\nabla u = \left(\dfrac{\partial u}{\partial x_1}, \dfrac{\partial u}{\partial x_2}, \ldots, \dfrac{\partial u}{\partial x_d}\right)$, and for a vector function $f(u(x)) = (f_1(u), f_2(u), \cdots, f_d(u))$, denoted $div(f(u)) = \sum_{i=1}^{d} \dfrac{\partial f_i(u)}{\partial x_i}$. To keep the following expression simple, we also denote $h_0' \triangleq h$, $h_i' = \dfrac{\partial h}{\partial x_i}$, for $i = 1, 2, \cdots, d$. According to the Gateauex differential definition(See [62]), we have for $\forall\ h \in L^2(R^d)$,

$$\begin{aligned}
& \delta E(u) = \lim_{\lambda \to 0} \frac{E(u + \lambda h) - E(u)}{\lambda} \\
= & \int_\Omega \lim_{\lambda \to 0} \frac{f(u + \lambda h, \nabla u + \lambda \nabla h) - f(u, \nabla u)}{\lambda} dx \\
= & \int_\Omega \sum_{i=0}^{d} f_i'(u, \nabla u) h_i' dx \\
= & \int_{\partial\Omega} \sum_{i=0}^{d} f_i'(u, \nabla u) h dx - \int_\Omega div(\nabla f(u, \nabla u)) h dx \\
\stackrel{*}{=} & -\int_\Omega div(\nabla f(u, \nabla u)) h dx
\end{aligned} \tag{2.4.1}$$

The last equation $\stackrel{*}{=}$ is established given that $\int_{\partial\Omega} \sum_{i=0}^{d} f_i'(u, \nabla u) h dx = 0$.

let

$$f(y_0, y_1, \cdots, y_d) = \frac{1}{2} \sum_{i,j=1}^{d} a_{ij} y_i y_j + \sum_{i=1}^{d} b_i y_0 y_i \tag{2.4.2}$$

It is clear that

$$f_0'(y_0, y_1, \cdots, y_d) = \sum_{i=1}^{d} b_i y_i \tag{2.4.3}$$

and for $k = 1, 2, \ldots, d$,

$$\begin{aligned} f_k^{'}(y_0, y_1, \cdots, y_d) &= \frac{1}{2}\left(\sum_{j=1}^{d}(a_{kj}y_j + a_{jk}y_j)\right) + b_k y_0 \\ &= \sum_{j=1}^{d} a_{kj}y_j + b_k y_0 \end{aligned} \tag{2.4.4}$$

thus

$$\begin{aligned} div(\nabla f(u, \nabla u)) &= \sum_{i=1}^{d} \frac{\partial f_i'(u, \nabla u)}{\partial x_i} \\ = \sum_{i,j=1}^{d} a_{ij}\frac{\partial^2 u}{\partial x_i x_j} &+ \sum_{i=1}^{d} b_i \frac{\partial u}{\partial x_i} \end{aligned} \tag{2.4.5}$$

since h is an arbitrary function, to obtain $\delta E(u)$, we see that we only need to have

$$div(\nabla f(u, \nabla u)) = 0,$$

which according to the steepest gradient descend method, the following PDE is required

$$\frac{\partial u(t, x)}{\partial t} = div(\nabla f(u, \nabla u)) = L_t u(t, x) \tag{2.4.6}$$

which proves Theorem 7. ■

Chapter 3

Nonlinear controlled diffusion

A nonlinear controlled diffusion[56, 28, 26, 71] is a different strategy to achieve a spatially varying target diffusion and offer a framework for a better understanding of nonlinear PDEs and corresponding diffusions. We first give a brief introduction to controlled diffusion, and subsequently apply the latter to obtain a different and insightful respective on nonlinear diffusion.

3.1 Definition of Nonlinear Controlled Diffusion

Definition 2. *If the drift and diffusion terms $b(t,x), \sigma(t,x)$ of SDE Eq. (2.2.3) are associated with a function $v(t,x)$, and are denoted by $b(t,x,v(t,x)), \sigma(t,x,v(t,x))$, $X_t^v = X_t$ is called a controlled diffusion, where X_t^v satisfies the SDE*

$$dX_t^v = b(t, X_t, v(t, X_t))dt + \sigma(t, X_t, v(t, X_t))dB_t. \tag{3.1.1}$$

where

$$L_t^v = \frac{1}{2}\sum_{i,j=1}^{d} a_{ij}(t,x,v(t,x))\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(t,x,v(t,x))\frac{\partial}{\partial x_i} \tag{3.1.2}$$

and $a(t,x,v(t,x)) = \sigma(t,x,v(t,x))\sigma(t,x,v(t,x))^T$.

Following Kolmogorov's backward and forward equations, we will have corresponding theorems if a number of conditions are satisfied. Here we only list some typical solutions of

the PDE with an infinitesimal operator as given in Eq. (3.1.2)

$$\begin{aligned} \frac{\partial U_t(x)}{\partial t} &= L_t^v U_t(x) \\ U_s(x) &= f(x) \quad \text{for some} \quad 0 \le s \le t, \end{aligned} \tag{3.1.3}$$

Its solution may be written as an expected value $U_t(x) = E_x\{f(X_t)\} = \int f(y)P(s,x,t,dy)$ [25]. Another backward equation has the following form:

$$\begin{aligned} \frac{\partial U_t(x)}{\partial t} + L_t^{v'} U_t(x) &= 0 \\ U_T(x) = f(x) \quad &\text{for some} \quad T \ge t, \end{aligned} \tag{3.1.4}$$

where $L_t^{v'}(\cdot)$ is another adjoint operator of L_t^v similar to that given in Eq. (2.2.13). The solution of this equation can again be expressed as $U_t(x) = E(f(\hat{X}_{st}(x)))$, namely, a probabilistic mean value of a reverse time process $\hat{X}_t, 0 \le t \le T$ where T is the fixed terminal/end time of the diffusion (or initial in the case of an inverse diffusion). Note that a backward diffusion can be viewed as a forward diffusion by merely selecting $t' = T - t$ as stated in Th. 6.

we need to, however, mention here that there are many properties of nonlinear controlled diffusions which remain as open problems. Our approach, here is to use the properties of linear diffusion as an inspiration for investigating some practical nonlinear problems whose numerical implementation is of primary concern.

3.2 Nonlinear Diffusion and PDE

As noted in the previous chapter, the equivalence between a Gaussian filter and heat equation-based evolution, led Witkin [95] to propose the following equation for filtering a noisy observation of a signal/image $f(\vec{x})$, where we hereafter denote $x \in \mathcal{R}^d$ as $\vec{x}$, namely $\vec{x} = (x_1, \cdots, x_d)$, to separate a vector in $\mathcal{R}^d$ from a scalar in $\mathcal{R}^1$,

$$\frac{\partial U_t(\vec{x})}{\partial t} = \Delta U_t(\vec{x}), \tag{3.2.1}$$

where $U_t(\vec{x})$ denotes the data at scale $t \in \mathcal{R}^+$, and $U_0(\vec{x}) = f(\vec{x})$ is the initial data, where $\vec{x} = x \in \mathcal{R}^1$ denotes a noisy signal $f(x)$ taking value in 1-D space, while $\vec{x} = (x_1, x_2) \in \mathcal{R}^2$

implies spatial coordinates of individual pixels of an image. "Δ" is the Laplacian operator (i.e., $\frac{\partial^2}{\partial x^2}$ when $d = 1$ or $\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$ when $d = 2$).[1] The function $U(\cdot, \cdot)$ at the finest scale ($t = 0$) is assumed to be comprised of a signal/image of interest and of a white Gaussian noise of variance σ^2.

Using the linear heat equation as their paradigm, Perona and Malik([75]) proposed to modify the evolution in Eq. (3.2.1) so as to achieve maximal smoothing in homogeneous regions of an image to eliminate noise, and minimal smoothing in high gradient regions to preserve edges. The proposed evolution equation which will be central to our development,[2] is written as,

$$\frac{\partial U_t(\vec{x})}{\partial t} = div\left(g\left(\mid \nabla\left(U_t(\vec{x})\right) \mid\right) \nabla U_t(\vec{x})\right), \tag{3.2.2}$$

where "div" represents the divergence operator, ∇ is the gradient operator, and $g(\cdot)$ is some measure of "edginess", thus a functional which modulates the strength of the diffusion according to the above paradigm (i.e., positive and monotonously decreasing with $g(0) = 1$). One possible choice is $g(v) = e^{-\frac{v^2}{K^2}}$ where K, a parameter to be judiciously chosen, determines the rate of decay and thus the extent of smoothing of $U_t(\vec{x})$ for a given gradient size. Because of space limitations, mathematical details as well as numerous other improvements on the P-M equation will not be discussed and deferred for instance to [89].

To solve Eq. (3.2.1), we use the stochastic interpretation, where the underlying particle motion is a Brownian motion, and for which, according to Th. 3 and Th. 6 of Chapter 1, we proceed to write

$$U_t(\vec{x}) = E_{0,\vec{x}}(f(\hat{X}_t)), \tag{3.2.3}$$

with $U_T(\vec{x}) = f(\vec{x})$. We also write

$$U_t(\vec{x}) = E(f(\hat{X}_{st}(\vec{x}))) \stackrel{\triangle}{=} E_{t,\vec{x}}(f(\hat{X}_{st})), \tag{3.2.4}$$

where $E_{t,\vec{x}}$ specifies that the inverse diffusion $\hat{X}_{st}$ is beginning from $\vec{x}$ at time t. namely, $U(t, \vec{x}) = f(\vec{x})$ is the initial data, and for a specific infinitesimal operator, such as L_t is a Laplacian operator, we have $\hat{X}_t = B_t$, this is due to the homogenous and symmetric properties of Brownian motion. However, for most processes that don't have these properties,

[1] The variable t in this context and throughout, represents scale instead of time.

[2] Note that many good techniques have since appeared, and to the best of our knowledge, all are prone to the same over-smoothing problem which is addressed herein.

Eq. (3.2.4) is an adaptive probability solution. In particular, we obtain one step transition as

$$\begin{aligned} U(t,\vec{x}) &= E_{t,\vec{x}}(f(\hat{X}_{st})) \\ &= \int E_{t-\tau,y}(f(\hat{X}_{s(t-\tau)}))p(\tau,\vec{x},\vec{y})d\vec{y} \\ &= \int U(t-\tau,\vec{y})P_\tau(\vec{y}|\vec{x})d\vec{y}. \end{aligned} \tag{3.2.5}$$

Note that the above diffusion $\hat{X}_t$ is a backward diffusion which, as mentioned in Chapter 1, is treated as a forward diffusion for ease of exposition and in the interest of clarity. The probability $P_\tau(\vec{y}|\vec{x})$ should then be interpreted as $P(\hat{X}_{t-\tau} = \vec{y}|\hat{X}_t = \vec{x})$ for a homogenous process and to emphasize the backward evolution in time/scale.

3.3 Nonlinear Diffusion on a Lattice

In light of the foregoing development, and for better insight and intuitive clarity, we find it useful to carry out most of the exposition and the analysis in a discrete setting and hence carry out the computation on a discrete lattice. Prior to delving into our formulation and interpretation of a Non-Linear (NL) diffusion, we present an illustrative example where the so-called controlled diffusion leads to a Markov Chain following an an optimization problem.

3.3.1 Discrete Approximation of Diffusion

As previously noted, our chief interest here is to propose a framework within which a stochastic interpretation of a diffusion (or more generally of the so-called scale space analysis) is achieved, and is in turn, instrumental in gaining insight. Towards that end and to further extend and possibly improve on existing techniques, we begin by discretizing the space as well as the scale/time variables.

Recall that a symmetric one-dimensional (1-D) random walk is well known to converge to a Brownian motion as $\tau \to 0$ and $\delta \to 0$, with τ, δ respectively denoting scale and distance discrete step size. A particle following such a trajectory will move on a 1-D lattice with probability 1/2 to the left or to the right, while on a 2-D plane, it will move to any of the

four nearest neighbors (east, west, north, south) with equal probability of 1/4. Formally, in 2-D space, we write the spatial variable $(x_{1i}, x_{2i}) = (x_1 + i\delta, x_2 + i\delta)$ with $i \in \mathbb{Z}$ and the scale $t_n = n\tau$ with $n \in \mathbb{N}$, we denote the one step transition probability of a particle from initial position (x_1^0, x_2^0) to (x_1, x_2) at the n^{th} scale step, by $p_n((x_1^0, x_2^0), (x_1, x_2))$. As a result, we obtain a standard form from Eq. (2.2.14), namely the probability of a particle being at (x_1, x_2) at scale/time $(n+1)^{st}$ step as $\tau \to 0$ and $\delta \to 0$,

Proposition 1. *The following discrete equation,*

$$\begin{aligned} p_{n+1}((x_1^0, x_2^0), (x_1, x_2)) &= \tfrac{1}{4}p_n((x_1^0, x_2^0), (x_1 - \delta, x_2) + \tfrac{1}{4}p_n((x_1^0, x_2^0), (x_1 + \delta, x_2)) \\ &\quad + \tfrac{1}{4}p_n((x_1^0, x_2^0), (x_1, x_2 - \delta)) + \tfrac{1}{4}p_n((x_1^0, x_2^0), (x_1, x_2 + \delta)) \end{aligned} \tag{3.3.1}$$

converges to

$$\frac{\partial p_t((x_1^0, x_2^0), (x_1, x_2))}{\partial t} = \frac{\partial^2 p_t((x_1^0, x_2^0), (x_1, x_2))}{\partial x_1^2} + \frac{\partial^2 p_t((x_1^0, x_2^0), (x_1, x_2))}{\partial x_2^2} \tag{3.3.2}$$

Proof: Subtracting $p_n((x_1^0, x_2^0), (x_1, x_2))$ from both sides of Eq. (3.3.1) and dividing it by τ, we obtain

$$\begin{aligned} &[p_{n+1}((x_1^0, x_2^0), (x_1, x_2)) - p_n((x_1^0, x_2^0), (x_1, x_2))]/\tau = \\ &\frac{1}{4\tau}\left[p_n((x_1^0, x_2^0), (x_1 - \delta, x_2)) - 2p_n((x_1^0, x_2^0), (x_1, x_2)) + \right. \\ &\left. p_n((x_1^0, x_2^0), (x_1 + \delta, x_2))\right] + \\ &\frac{1}{4\tau}\left[p_n((x_1^0, x_2^0), (x_1, x_2 - \delta)) - 2p_n((x_1^0, x_2^0), (x_1, x_2)) + \right. \\ &\left. p_n((x_1^0, x_2^0), (x_1, x_2 + \delta))\right] \end{aligned}$$

which upon letting $\tau = \delta^2/4$ and $\delta \to 0$, concludes the proof. ∎

With numerical implementation of a linear diffusion in hand, we proceed to consider a 1-D Brownian motion on a compact interval. By defining a reflecting wall on this interval, we make the resulting Markov Chain(MC) aperiodic and recurrent with a solution to $\Delta U(\vec{x}) =$

$0, U_T(\vec{x}) = f(\vec{x})$ taking the form $U(\vec{x}) = E_{\vec{x}}(f(X_T))$. This also implies a discrete solution $U(\vec{x}) = \sum_i p_i f(x_i)$ where p_i is the probability of a particle to be in state i as t grows large. The independence of the solution $U(\vec{x})$ of the initial state, implies its convergence to some mean value of of $f(\vec{x})$ as $\vec{x}$ is averaged over all possible paths. This in a sense provides an intuitive justification for the convergence of a heat equation to a constant. A case in point arises when we are faced with a deterministic diffusion X_t, which is alternatively expressed as $dX_t = (1 \;\; 0)dt$. The corresponding solution obtained from Eq. (3.1.3) is $U(\vec{x}) = f(x_1^0 + t, x_2^0)$ where $\vec{x_0} = (x_1^0, x_2^0)$ is the initial state, clearly non-constant as expected.

3.3.2 Finite Markov Chain Example

Let a finite Markov chain [27] as in Fig. 3.1 with three possible states α, β, γ with respective costs 1, 2, 3. Our goal is to select a "best" strategy (minimum cost) for a particle to make a two-step transition from a state, say α. The corresponding transition probabilities are obtained from the following sets of strategies:

$$\begin{aligned} G_\alpha &= \{(\frac{1}{2}, 0, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2}, 0)\} \\ G_\beta &= \{(\frac{1}{3}, 0, \frac{2}{3}), (\frac{3}{4}, \frac{1}{4}, 0)\} \\ G_\gamma &= \{(1, 0, 0), (\frac{1}{2}, 0, \frac{1}{2}), (\frac{1}{3}, 0, \frac{2}{3})\}. \end{aligned}$$

For state α for instance, two choices are possible:

- the first strategy has it move to state γ with probability $\frac{1}{2}$, and remain stationary with probability $\frac{1}{2}$,
- the second lets it move to state β with probability $\frac{1}{2}$ and remain stationary with probability $\frac{1}{2}$.

Taking into account the respective costs as well as the transition probabilities, an optimal strategy for α is determined to be $(\frac{1}{2}, \frac{1}{2}, 0)$ with a minimum cost of $\frac{3}{2}$, while in state β a cost of $\frac{5}{4}$ with the strategy $(\frac{3}{4}, \frac{1}{4}, 0)$, and in state γ we obtain a cost of 1 with strategy $(1, 0, 0)$. A second step transition may be similarly found, with respective costs for α, β, γ, of

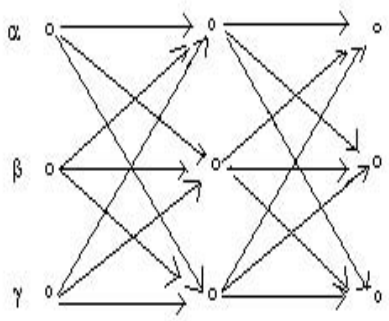

Figure 3.1: Finite Controlled Markov Chain modelling.

$\frac{5}{4}, \frac{7}{6}, \frac{7}{6}$ resulting from $(\frac{1}{2}, 0, \frac{1}{2}), (\frac{1}{3}, 0, \frac{2}{3}), (\frac{1}{3}, 0, \frac{2}{3})$ respectively. This process may be continued indefinitely.

Note that more complex strategies are possible and may be constructed for the intermediate steps, e.g., variable strategies along the steps, etc.. When, on the other hand, a given strategy set only depends on the previous state, it is referred to as a Markov strategy. It is also clear from the foregoing example that the resulting process, by way of its transition strategy, influences the overall mean value.

This example provides a motivation to pursue such an approach of controlling diffusion(drift and diffusion coefficients) and sufficient evidence for it to be a promising and systematic way of addressing problems in image enhancement/segmentation, and shedding light on the current outstanding problems in nonlinear diffusion.

3.3.3 Discrete Time/Scale Evolution

By discretizing $\vec{x}$ and t, we can account for a reverse time evolution by relabelling time "$t-\tau$" by 1 (or τ) and "$t-n\tau$" by n (or $n\tau$), hence making a backward diffusion equation look more like a forward diffusion. We denote by $U_n(\vec{x})$ the value of the solution at time

step $n\tau$ and location/state $\vec{x}$. Eq. (3.2.5) can then be written in the form,

$$U_{n+1}(\vec{x}) = \sum_{\vec{y}} U_n(\vec{y}) P_{n+1}(\vec{y}|\vec{x}). \tag{3.3.3}$$

where $P_{n+1}(\vec{y}|\vec{x})$ denotes the probability for a particle to move to $\vec{y}$ at step $n+1$ with an initial position $\vec{x}$ at step n. Since the solution to Eq. (3.3.2) is a Gaussian transition density function, it characterizes the evolution of a particle along a Brownian trajectory starting at $\vec{x_0}$ and time t. Using the fact that a limiting process of a random walk is a Brownian motion, we may compute the solution to Eq. (3.3.3) at any desired discrete time/scale. At the first time step τ and for a 1-D case, we can write

$$U_1(x) = \frac{1}{2} f(x-\delta) + \frac{1}{2} f(x+\delta), \tag{3.3.4}$$

while in a 2-D scenario, we have

$$U_1(x_1, x_2) = \frac{1}{4} f(x_1-\delta, x_2) + \frac{1}{4} f(x_1+\delta, x_2) + \frac{1}{4} f(x_1, x_2-\delta) + \frac{1}{4} f(x_1, x_2+\delta), \tag{3.3.5}$$

both of which are the result of an averaging process. More generally, we can respectively write the 1-D and 2-D solutions to the linear heat equation as discrete expectations

$$U_{n+1}(x) = \frac{1}{2} U_n(x-\delta) + \frac{1}{2} U_n(x+\delta), \tag{3.3.6}$$

$$\begin{aligned} U_{n+1}(x_1, x_2) = {} & \frac{1}{4} U_n(x_1-\delta, x_2) + \frac{1}{4} U_n(x_1+\delta, x_2) + \frac{1}{4} U_n(x_1, x_2-\delta) + \\ & \frac{1}{4} U_n(x_1, x_2+\delta) \end{aligned} \tag{3.3.7}$$

Due to the underlying random walker moving to its neighbor with probability 1/2 in 1-D (and to its four nearest neighbors with probability 1/4 in 2-D), it is clear that the linear evolution will indiscriminately smooth away sharp features along with the noise.

3.3.4 A Stochastic View of Perona-Malik Equation

As noted earlier, a linear stochastic differential equation leads to a linear diffusion by way of a Laplacian as its corresponding infinitesimal generator. Using this development as an

inspiration together with its discrete stochastic formulation and interpretation, we proceed in an analogous manner to rewrite the P-M equation to be interpreted as a particle-based diffusion.

Proposition 2. *Based on a particle system interpretation, P-M equation may be rewritten as*

$$\begin{aligned} U_{n+1}(x) &= p^{n+1}(x, x+\delta)U_n(x+\delta) + p^{n+1}(x, x-\delta)U_n(x-\delta) + \\ &\quad [1 - p^{n+1}(x, x+\delta) + p^{n+1}(x, x-\delta)]U_n(x). \end{aligned} \tag{3.3.8}$$

Proof: The proof follows immediately from discretizing Eq. (5.2.1) and rewriting $1/2g\,(\mid U_n(x \pm \delta) - U_n(x) \mid) = p^{n+1}(x, x \pm \delta) = p^{n+1}(\xi_{n+1} = x \pm \delta \mid \xi_n = x)$ to denote the transition probability of a Markov chain $\left\{\xi_{(\cdot)}\right\}$ to move from state x to state $x \pm \delta$.

A similar expression for a 2-D signal (image) may be written as

$$\begin{aligned} & U_{n+1}(x_1, x_2) \\ = \; & p_S^{n+1}(x_1, x_2)U_n(x_1 + \delta, x_2) + p_N^{n+1}(x_1, x_2)U_n(x_1 - \delta, x_2) \\ & + p_E^{n+1}(x_1, x_2)U_n(x_1, x_2 + \delta) + p_W^{n+1}(x_1, x_2)U_n(x_1, x_2 - \delta) \\ + \; & [1 - p_S^{n+1}(x_1, x_2) - p_N^{n+1}(x_1, x_2) - p_E^{n+1}(x_1, x_2) - p_W^{n+1}(x_1, x_2)]U_n(x_1, x_2), \end{aligned} \tag{3.3.9}$$

where

$$p_S^{n+1}(x_1, x_2) = p_N^{n+1}(x_1, x_2) = p_E^{n+1}(x_1, x_2) = p_W^{n+1}(x_1, x_2) = \frac{1}{4} g\,(\mid \nabla U \mid)$$

and

$$\mid \nabla U \mid = \sqrt{\nabla U_1^2 + \nabla U_2^2 + \nabla U_3^2 + \nabla U_4^2} \tag{3.3.10}$$

$$\begin{aligned} \nabla U_1 &= U_n(x_1 + \delta, x_2) - U_n(x_1, x_2) \\ \nabla U_2 &= U_n(x_1 - \delta, x_2) - U_n(x_1, x_2) \\ \nabla U_3 &= U_n(x_1, x_2 + \delta) - U_n(x_1, x_2) \\ \nabla U_4 &= U_n(x_1, x_2 - \delta) - U_n(x_1, x_2) \end{aligned}$$

The probabilities $p_S^{n+1}(x_1, x_2)$ (resp. $p_N^{n+1}(x_1, x_2)$, $p_E^{n+1}(x_1, x_2)$, $p_W^{n+1}(x_1, x_2)$, $p_S^{n+1}(x_1, x_2)$) represent the transition probabilities of the underlying Markov chain ξ_n, i.e., $p_S^{n+1}(x_1, x_2) = p_S(\xi_{n+1} = (x_1 + \delta, x_2) \mid \xi_n = (x_1, x_2))$ (similar expressions for other direction transition probabilities). which says that activities of diffusion in 4 directions are uniform in the same location but different and decided by the local gradient measure in different locations, the variational functional for d(in special, $d = 2$) that corresponded to the P-M equation is given as

$$E(u) = \frac{1}{2} \int_{R^d} \left(1 - \exp\{- \mid \nabla u \mid^2 /K\}\right) dx \tag{3.3.11}$$

The widely used implementation of the PM-algorithm whose interpretation herein is given below, is simpler and better adapted to image processing.

$$\begin{aligned} p_S^{n+1}(x_1, x_2) &= \frac{1}{4} g\left(\mid \nabla U_1 \mid\right), \\ p_N^{n+1}(x_1, x_2) &= \frac{1}{4} g\left(\mid \nabla U_2 \mid\right), \\ p_E^{n+1}(x_1, x_2) &= \frac{1}{4} g\left(\mid \nabla U_3 \mid\right), \\ p_W^{n+1}(x_1, x_2) &= \frac{1}{4} g\left(\mid \nabla U_4 \mid\right) \end{aligned}$$

These equations are intuitively appealing, in that the random walk of a particle (or pixel) (or the diffusion) ξ_n takes place, in each direction, according to the prevailing one sided gradient at position (x_1, x_2) in any of the four directions. At time step $n + 1$, a south (resp. north, east, west) moving walk takes place with probability $p_S^{n+1}(x_1, x_2)$ (resp. $p_N^{n+1}(x_1, x_2)$, $p_E^{n+1}(x_1, x_2), p_W^{n+1}(x_1, x_2)$), and the particle remains in place with probability $p_0^{n+1}(x) = 1 - p_S^{n+1}(x_1, x_2) - p_N^{n+1}(x_1, x_2) - p_E^{n+1}(x_1, x_2) - p_W^{n+1}(x_1, x_2)$.

The variational formulation which yields the above measures in 2-D is

$$E(u) = \frac{1}{2} \int_{R^d} \sum_{i=1}^{d} \left(1 - \exp\left\{- \left(\frac{\partial u}{\partial x_i}\right)^2 /K\right\}\right) dx \tag{3.3.12}$$

Strictly speaking, This expression is different from that of P-M in Eq. (5.2.1) as only the gradient of each individual component affects the transition. According to Eq. (2.4.1), we have $f(x_1, x_2, \cdots, x_d) = \frac{1}{2} \sum_{i=1}^{d} (1 - e^{-x_i^2/K})$, which, when used for $E(u)$, yields the following

Gateauex difference,

$$\begin{aligned} \delta E(u) &= \int_\Omega \lim_{\lambda\to 0} \frac{f(\nabla u + \lambda\nabla h) - f(\nabla u)}{\lambda} dx \\ &= -\int_\Omega div(\nabla f(\nabla u))h dx \end{aligned} \tag{3.3.13}$$

using the derivative

$$f_i'(x_1, \cdots, x_d) = \frac{x_i}{K} e^{-x_i^2/K} \tag{3.3.14}$$

and neglecting the constant coefficient K on the right hand side of the formula, we have

$$\begin{aligned} div(\nabla f(u, \nabla u)) &= \sum_{i=1}^{d} \frac{\partial f_i'(\nabla u)}{\partial x_i} \\ &= \sum_{i=1}^{d} \frac{\partial}{\partial x_i} \left(\frac{\partial u}{\partial x_i} e^{-(\frac{\partial u}{\partial x_i})^2/K} \right). \end{aligned} \tag{3.3.15}$$

The corresponding steepest gradient descent is

$$\frac{\partial u}{\partial t} = \sum_{i=1}^{d} \frac{\partial}{\partial x_i} \left(\frac{\partial u}{\partial x_i} \exp\left\{ -\left(\frac{\partial u}{\partial x_i} \right)^2 / K \right\} \right) \tag{3.3.16}$$

It is clear here that the transition probability of such a random walk is determined by the gradient, inducing the desired control on the diffusion. This is in sharp contrast to the linear diffusion where the random walk invariably takes place with a constant probability of 1/4. Note that while the derivation of an exact SDE corresponding to P-M equation as an infinitesimal generator, is interesting in it and of itself, a more complex system of particles which is of little relevance to our stated goal in this paper, is required.

3.4 Two Sided Gradient-Driven Diffusion

As discussed in Section 2, at each scale of our analysis, the mean value of the process $U(\cdot, \cdot)$ is evaluated as a result of a non-homogeneous random walk with the transition probability controlled by the underlying process at the previous scale. In addition, and to avoid potential stability problems, we ensure that the probability of a jump of a particle (pixel) farther

than an immediate neighbor is zero, which effectively emulates a continuous diffusion. Furthermore, we ensure that there always be a one step transition of a particle to its neighbors to avoid a slowdown in convergence due to likely stationary states [56]. We thus adopt this paradigm to construct a non-homogeneous Markov chain whose transition probabilities are based on the current particle states and their functional value. This results in a set of consecutive transition steps through scales, each in a sense, defining a new random process with a new probability transition.

While the goal in signal/image processing is to maximally smooth out the noise, we are also keen on achieving a solution that is as faithful as possible to the initial underlying signal. To thus help better localize the homogeneous regions together with their boundaries, we use in our transition dynamics a bidirectional gradient-based "probability measure". (subgradient in continuous space). Using the Szökefalvi-Nagy's inequality[72], to optimize the gradient energy (to delineate regions), we have to minimize the following energy expression,

$$\begin{aligned} & \mathcal{E}(U_{n+1}) = \sum_x \mathcal{E}(U_{n+1}(x)) \\ = & \sum_x [(U_n(x+\delta) - U_n(x))(U_{n+1}(x) - U_n(x-\delta))]^2 \\ + & [(U_n(x-\delta) - U_n(x))(U_{n+1}(x) - U_n(x+\delta))]^2 \end{aligned} \tag{3.4.1}$$

where $U_{n+1}(x)$, assumed to result from Eq. (3.2.5) is written as,

$$U_{n+1}(x) = P^{n+1}(x, x-\delta)U_n(x-\delta) + P^{n+1}(x, x+\delta)U_n(x+\delta) \tag{3.4.2}$$

with $P^{n+1}(x, x-\delta) + P^{n+1}(x, x+\delta) = 1$. Minimizing Eq. (3.4.1) entails an appropriate choice of a probability measure as follows,

Theorem 8. *The transition probability solving Eq. (3.4.1) is given by* $P^{n+1}(x, x-\delta) = P\{\xi_{n+1} = x - \delta | \xi_n = x\}$ *with*

$$P^{n+1}(x, x-\delta) = \frac{\mid U_n(x+\delta) - U_n(x) \mid^2}{\mid U_n(x-\delta) - U_n(x) \mid^2 + \mid U_n(x+\delta) - U_n(x) \mid^2}, \tag{3.4.3}$$

where $U_{n+1}(x)$ *satisfies Eq. (3.4.2).*

(See Appendix A for a proof). ■

For a 2-D image, we denote the transition probability by $p_S^{n+1}(x_1, x_2) = P\{\xi_{n+1} = (x_1 + \delta, x_2)|\xi_n = (x_1, x_2)\}$ (similarly for other probabilities) and obtain the following expression for the transition probability

$$p_S^{n+1}(x_1, x_2) = \frac{\mathcal{N}}{\mathcal{S} + \mathcal{N} + \mathcal{E} + \mathcal{W}}, \tag{3.4.4}$$

where

$$\begin{aligned} \mathcal{N} &= | U_n(x_1 - \delta, x_2) - U_n(x_1, x_2) |^2, \\ \mathcal{S} &= | U_n(x_1 + \delta, x_2) - U_n(x_1, x_2) |^2, \\ \mathcal{E} &= | U_n(x_1, x_2 + \delta) - U_n(x_1, x_2) |^2, \\ \mathcal{W} &= | U_n(x_1, x_2 - \delta) - U_n(x_1, x_2) |^2 . \end{aligned}$$

Using the above transition probability, our newly proposed diffusion is written as

$$\begin{aligned} U_{n+1}(x_1, x_2) &= U_n(x_1 + \delta, x_2)p_S^{n+1}(x_1, x_2) + U_n(x_1 - \delta, x_2)p_N^{n+1}(x_1, x_2) \\ & \quad U_n(x_1, x_2 + \delta)p_E^{n+1}(x_1, x_2) + U_n(x_1, x_2 - \delta)p_W^{n+1}(x_1, x_2) \end{aligned} \tag{3.4.5}$$

3.5 Discussion

As noted in the previous section, the P-M diffusion is driven by a one-sided gradient at any position x, which implies that weak smoothing takes place in the presence of a relatively high gradient, even if the latter is caused by noise. On the other hand, when we consider a two sided gradient at a position (x_1, x_2), we are better able to identify a noise-induced high gradient at that position, as it is likely to register a high value on both sides of the pixel (x_1, x_2) under consideration. This would then allow us to discriminate between a "true" high gradient and that caused by noise. This is in contrast to the P-M filter which relies on a one-sided gradient which calls for a no transition state in the Markov chain, thus preserving the prevailing singularity. Note that this technical difficulty of the P-M equation may further be compounded in that the transition policy in the Markov chain may eliminate true edges; since at a position $\vec{x}$ close to the leading edge of a signal, and if at position $(x_1 + \delta, x_2)$, $U(\cdot, x_1, x_2)$ is close to $U(\cdot, x_1 + \delta, x_2)$, the probability of transition is finite, and

the smoothing of this leading edge takes place. This scenario is a *zero measure* event when a two sided gradient-based transition probability is used in the policy.

Using a Markov chain $\{\xi_n\}$, we can thus model these dynamics via transition probabilities which, as we mentioned, may be specified in terms of the ratio of a bidirectional gradient. We should note that in the cases where the sub-gradients are very small, and hence little significance can be attached to their ratio, the motion of the particle is based on a symmetric random walk, i.e., with a 1/4 probability it moves to the four nearest neighbors, hence leading to a *linear heat-like* equation.

When on the other hand, we use a two-sided gradient, a discontinuity arises and the two gradients are very different, at least one of them grows large which blocks the diffusion in the other direction. It can also be seen from the transition probability expressions that the diffusion is driven to zero when a plateau is reached, i.e., a *stable/fixed point* at a staircase function [49].

3.6 Experimental Results

Our goal in this section is to substantiate the results that we have established in the previous sections. The stabilization of the proposed diffusion at staircase functions together with the subsequent denoising and segmentation effects are first demonstrated by running a noise free signal/image which remains unaffected by the diffusion as displayed in Figure 3.3.

To establish a basis for performance comparison with the P-M equation which, recall, was the source of inspiration for our proposed technique, we run experiments where both visual as well as quantitative assessments are inferred. The denoising performance can be evaluated visually as shown in Figures (3.4, 3.5, 3.6). Figure 3.4 demonstrates a segmentation/denoising of an infra-red real image of a boat, run to similar time/scale for both P-M and our proposed technique. This class of images is well known for posing great challenges to simple gradient-based segmentation and/or linear filtering, and this is for the most part due to their impulsive nature. The results of such approaches usually results in visually unpleasant and quantitatively inaccurate results if at all. In Fig. 3.6, the potential for a

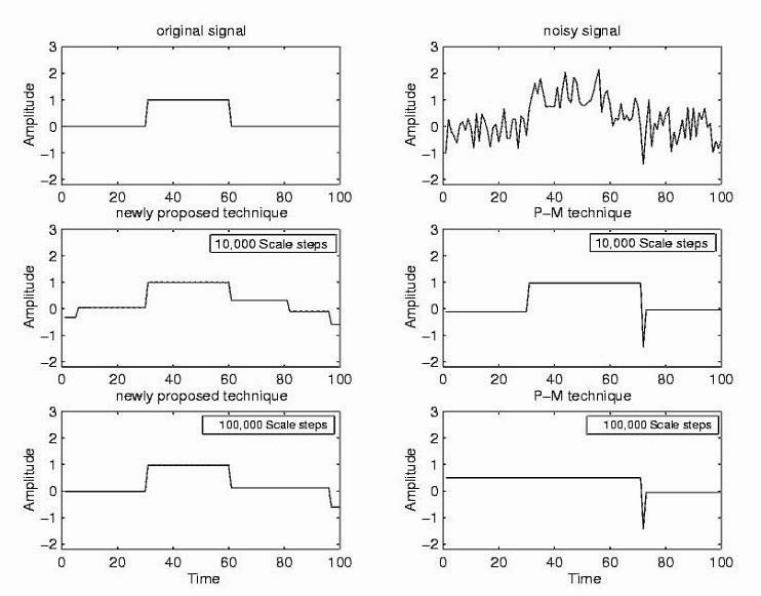

Figure 3.2: Noisy signal filtered by our random walk algorithm and PM algorithm.

complete diffusion (for a non-optimal choice of the threshold parameter in the P-M equation) is demonstrated whereas and as shown above, the new approach will stabilize with no parameter adjustment. For a well known stopping time and well chosen parameter, P-M approach performs quite well. This may, however, turn out to be a limitation for a number of real applications, as it is generally not known what the image consists of.

In Fig. 3.7 an enhancement/deblurring-like effect using our algorithm is also demonstrated for a checker board. The cost of doing away with the explicit knowledge of the stopping time for our approach, is the arising block effects which, although common to many existing techniques, remain a drawback. Although this may be fine tuned away for pure denoising purposes, we discuss in the next chapter techniques which specifically address such a shortcoming. In Fig. 3.7, a de-blurring example is shown, demonstrating the capacity of the algorithm to enhance edges and again to stabilize at staircase functions. In Figure(3.8),(3.9) we demonstrate denoising and segmentation results of images, which can serve as a comparison of our new algorithm and the P-M algorithm.

For establishing a more quantitative measure of performance we use the figures in Figure 3.10. A pixel deviation is computed and an error rate is defined as an unmatched segment(in the meaning of region segment) between filtered image(or noisy image) and

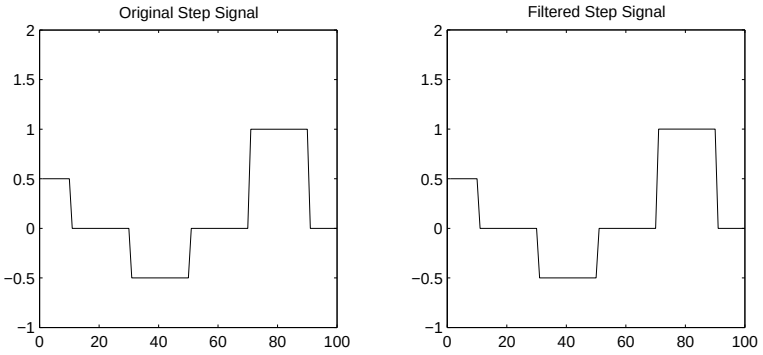

Figure 3.3: Stable signal remains unchanged following proposed nonlinear diffusion.

clear image. the error rate curve which is consistent with our visual assessment is displayed in Fig. 3.11.

3.7 Conclusion

The proposed stochastic interpretation together with its link to controlled diffusion are shown to not only explicate existing techniques and their limitations, but to also provide sufficient insight to develop other novel physically and geometrically driven methodologies. We have also succeeded in resolving in part, a well known and long standing problem of unknown stopping criterion.

3.8 Appendix A

Proof: We first proceed to re-express $\mathcal{E}(U_{n+1})$ in terms of Eq. (3.4.2) and $P_{n+1}(x, x+\delta) = 1 - P_{n+1}(x, x-\delta)$. By subsequently differentiating with respect to $P_n(x, x-\delta)$ and bearing in mind that a two sided-gradient is used, we have the following equation

$$\begin{aligned} & \partial(\mathcal{E}(U_{n+1}))/\partial P_{n+1}(x, x-\delta) \\ = \; & 2[(U_n(x+\delta) - U_n(x))^2 (U_{n+1}(x) - U_n(x-\delta))](\partial U_{n+1}(x))/\partial P_{n+1}(x, x-\delta) \\ + \; & 2[(U_n(x-\delta) - U_n(x))^2 (U_{n+1}(x) - U_n(x+\delta))](\partial U_{n+1}(x))/\partial P_{n+1}(x, x-\delta) \end{aligned} \tag{3.8.1}$$

Setting $\partial(\mathcal{E}(U_{n+1}))/\partial P_{n+1}(x, x-\delta) = 0$, and assuming a non-degenerate case of $\partial U_{n+1}(x)/\partial P_{n+1}(x, x-\delta) \neq 0$, the optimal transition probability at the n-th step implies

$$\begin{aligned} & [(U_n(x+\delta) - U_n(x))^2 + (U_n(x-\delta) - U_n(x))^2] U_{n+1}(x) \\ = \; & (U_n(x+\delta) - U_n(x))^2 U_n(x-\delta) + (U_n(x-\delta) - U_n(x))^2 U_n(x+\delta) \end{aligned} \tag{3.8.2}$$

where the replacement of $U_{n+1}(x)$ with Eq. (3.4.2) will reduce to Eq. (3.4.3). Note that if $\partial U_{n+1}(x)/\partial P_{n+1}(x, x-\delta) = 0$, we can see that a left sided-gradient is equal to the right sided-gradient resulting in an optimal choice of probability of 1/2. ■

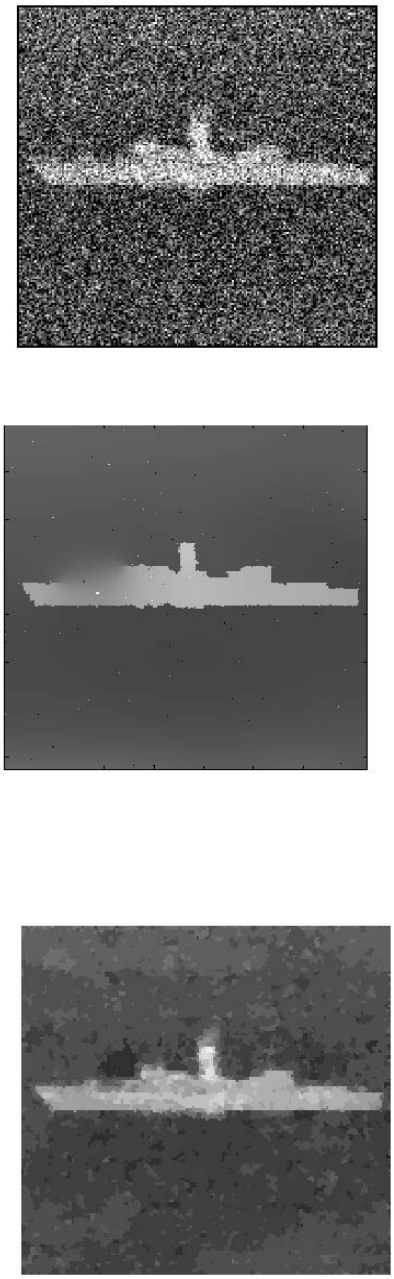

Figure 3.4: A noisy image together with its enhanced copy by the proposed algorithm and by the P-M method best result.

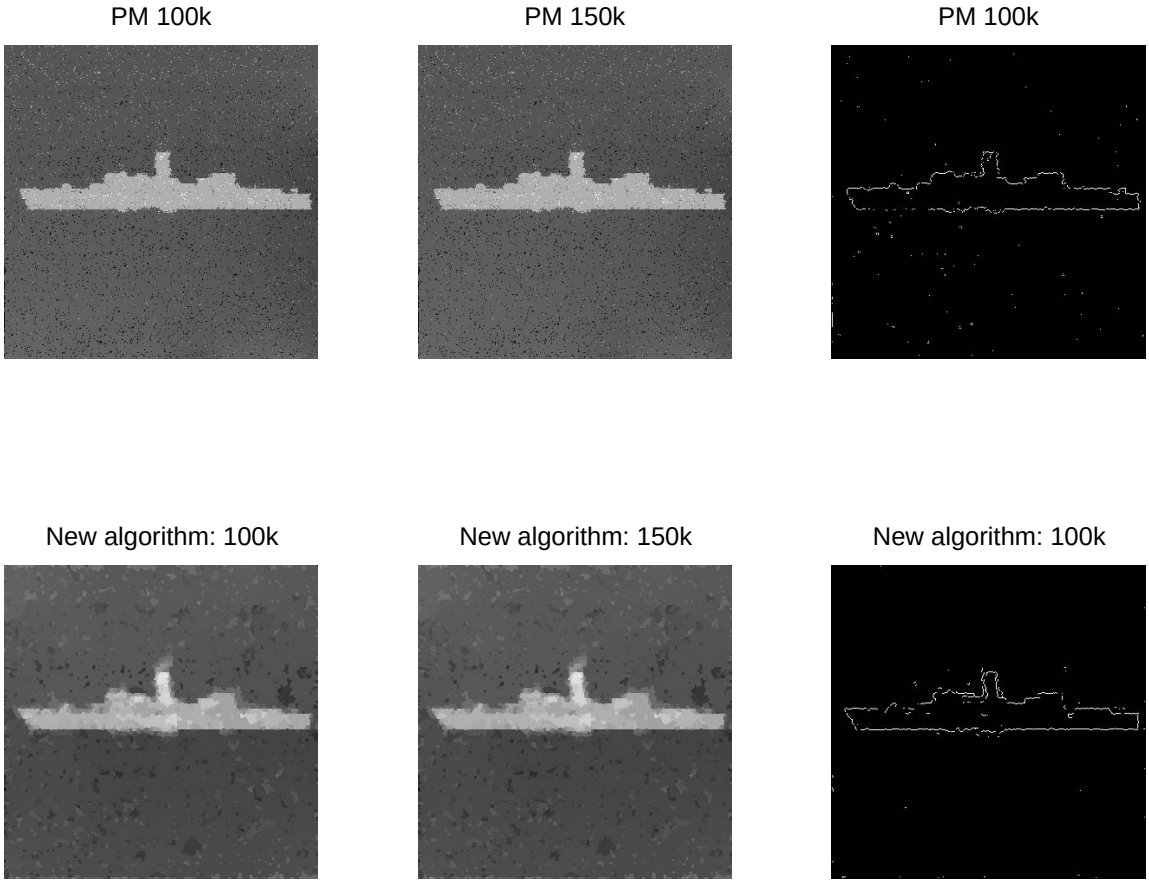

Figure 3.5: Complete Smoothing vs Stability.

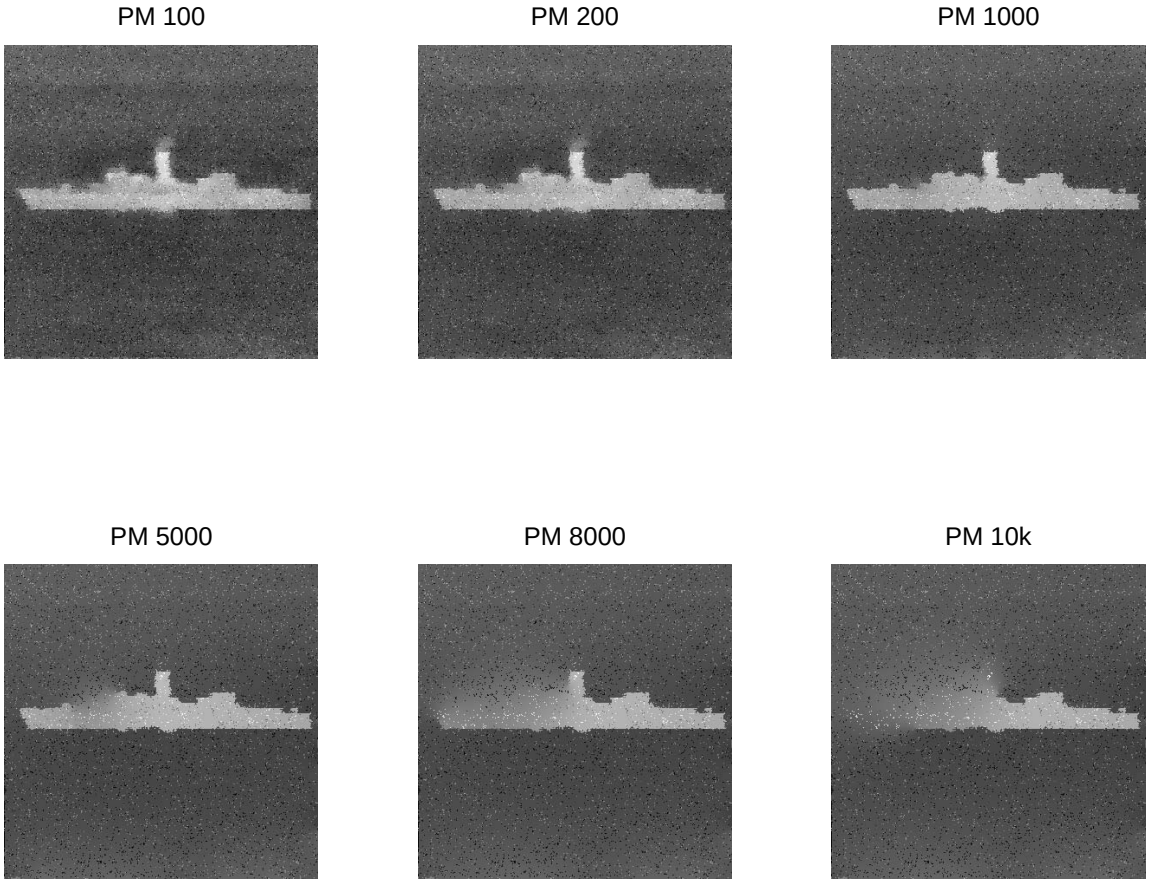

Figure 3.6: PM algorithm.

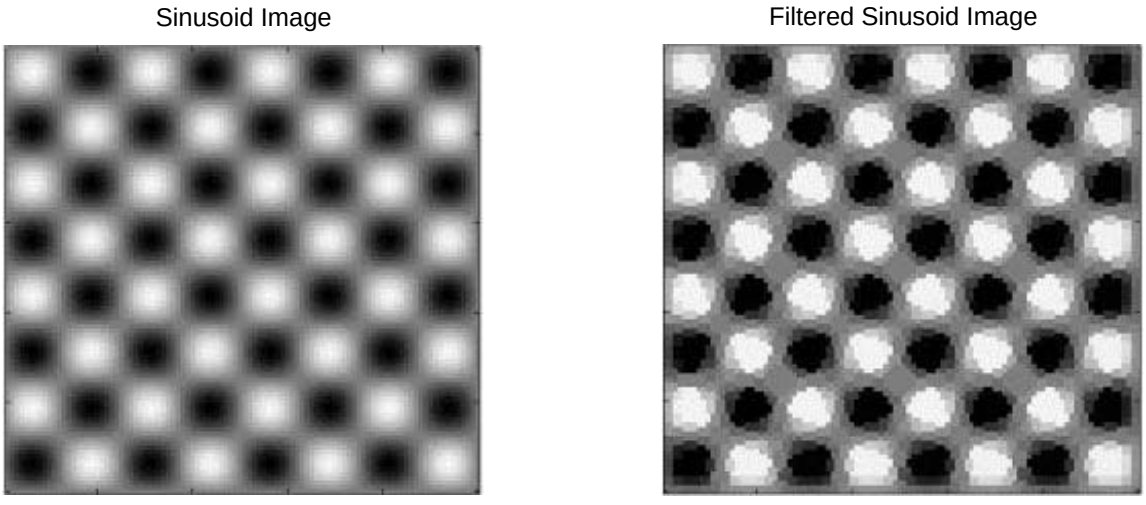

Figure 3.7: Checker Board Enhancement.

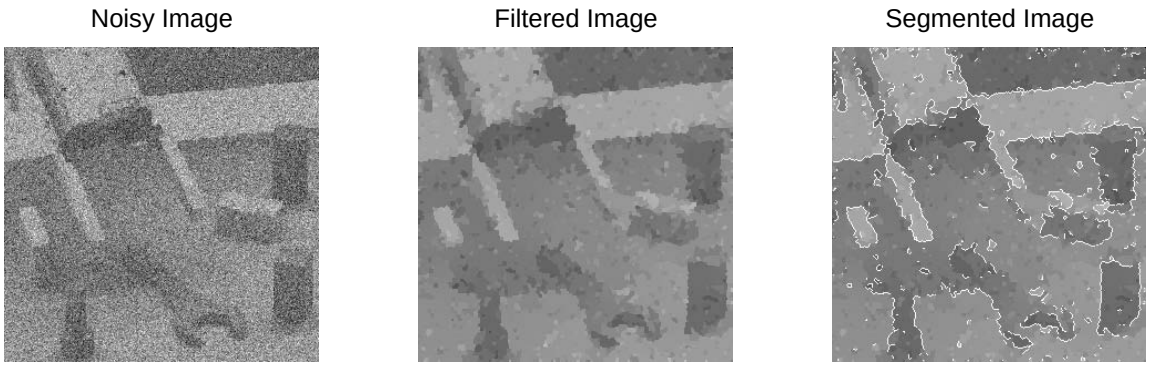

Figure 3.8: tools segmentation.

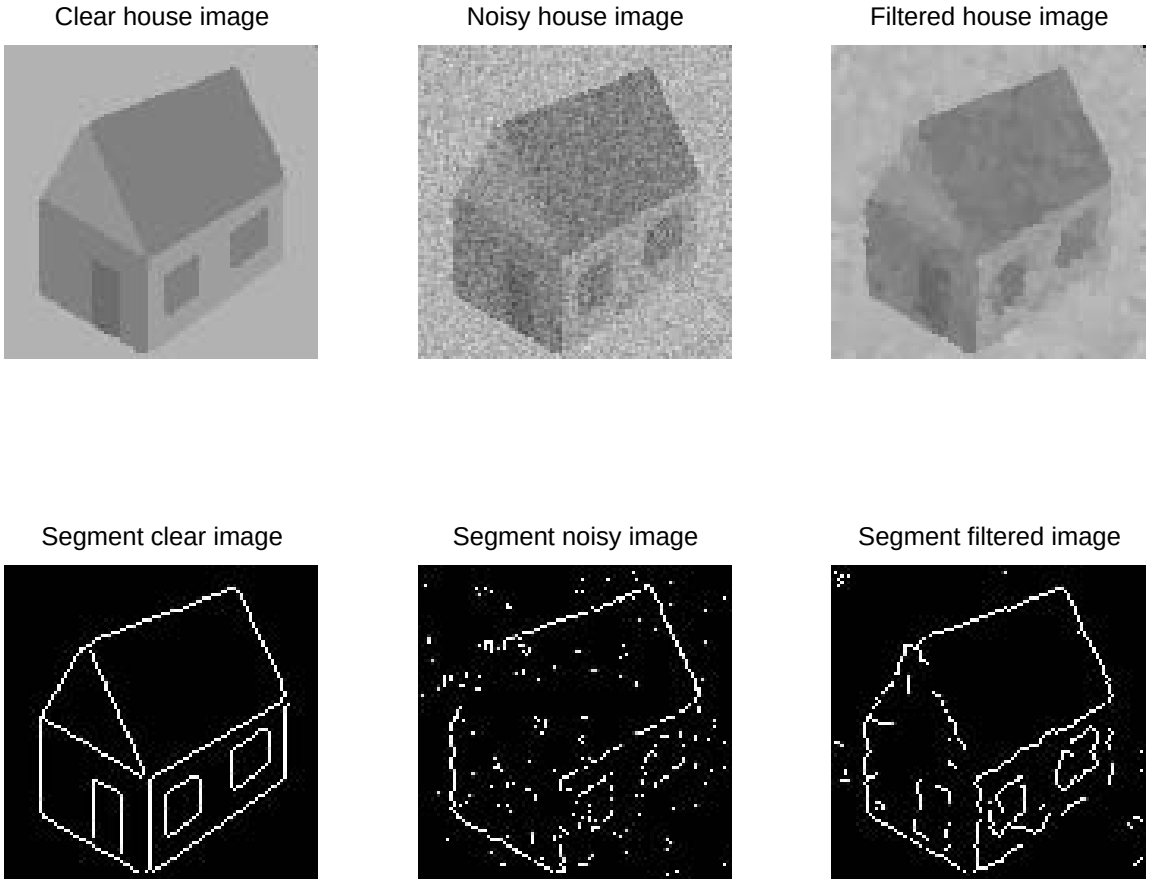

Figure 3.9: House segmentation.

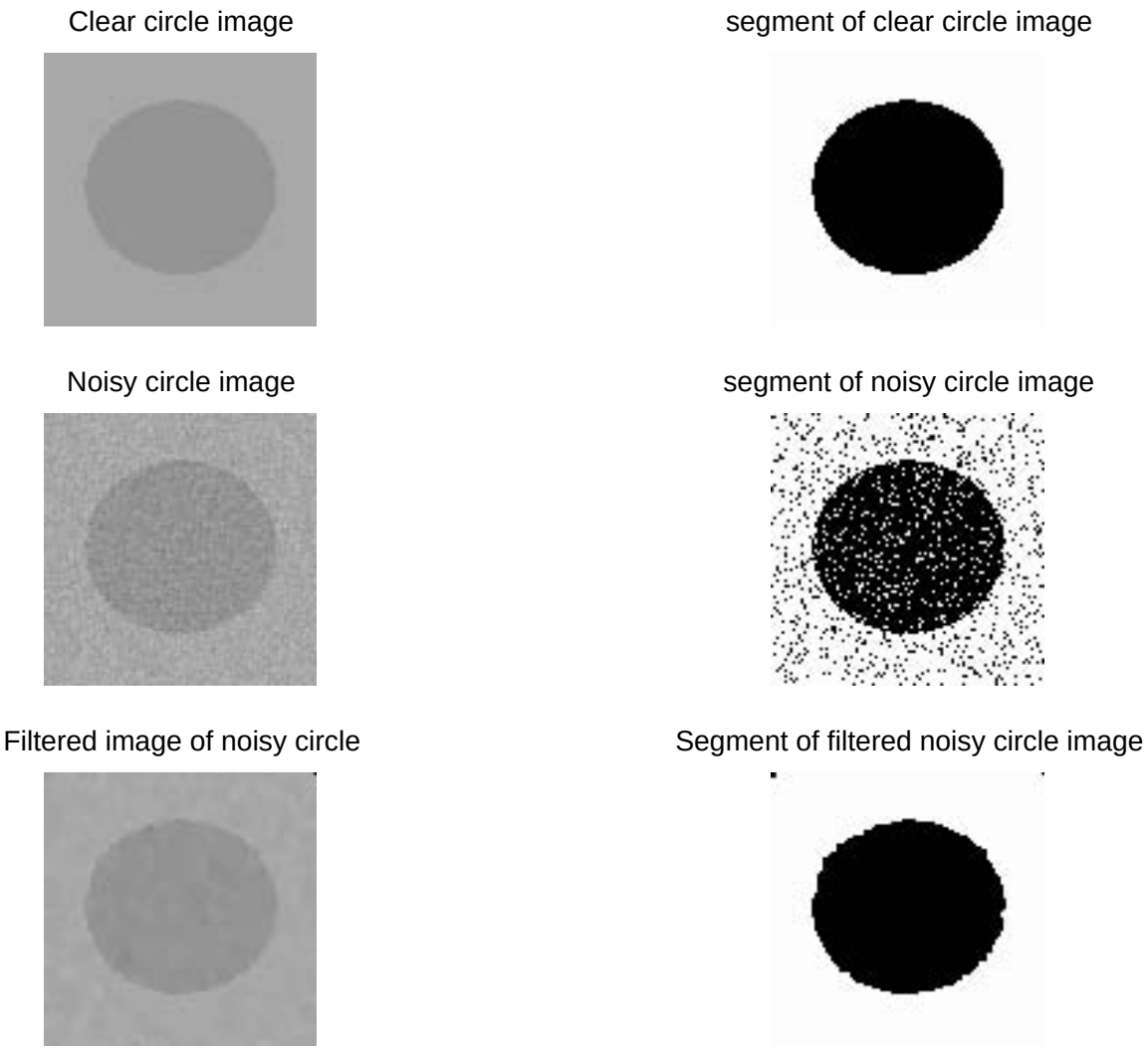

Figure 3.10: Circle segmentation.

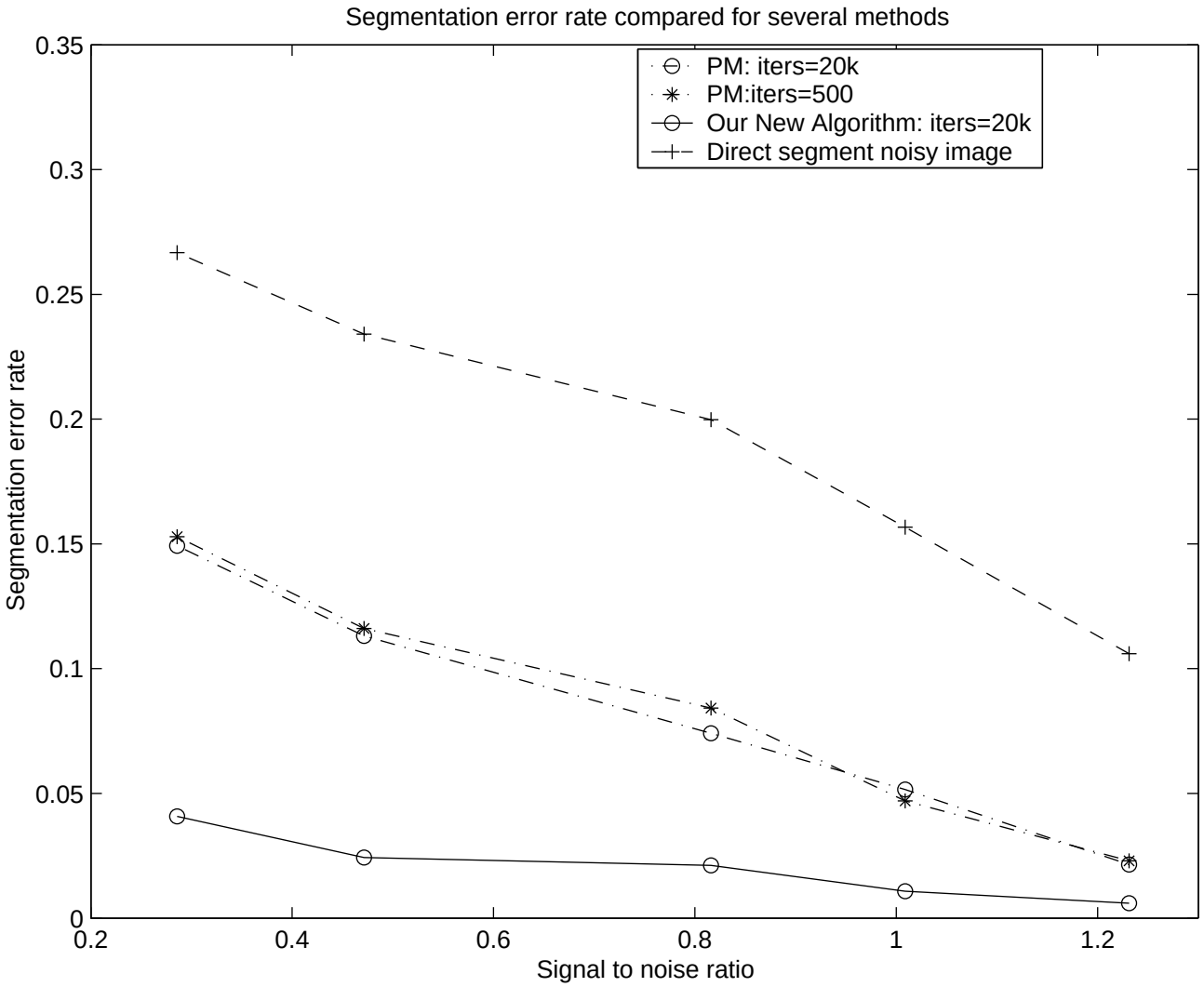

Figure 3.11: Error rate in circle segmentation for different SNR scenarios.

Chapter 4

Multiscale Wavelet space

Wavelet as a local signal analysis tool plays a very important role in obtaining detail information about signal structures. In this chapter, we introduce some basic facts about wavelets that would be useful for next chapter. Further details may be found in [66].

4.1 Definition of a Wavelet

Let $L^2(\mathcal{R})$ be a space of functions $f(x),\ x \in \mathcal{R}$, such that

$$||f||_2 = \sqrt{\int |f(x)|^2 dx} < +\infty \tag{4.1.1}$$

where the integral $||f||_2$ defines a norm of $L^2(\mathcal{R})$, and is also denoted by $||f||$ in this thesis.

A wavelet, also known as a mother wavelet, is a function $\psi(x) \in L^2(\mathcal{R})$ that is centered in the neighborhood of x with $||\psi|| = 1$ and zero mean:

$$\int_{-\infty}^{+\infty} \psi(x)dx = 0 \tag{4.1.2}$$

In the spectral domain, we denote by $\hat{\psi}(\omega)$ the Fourier transform of $\psi(x)$, and by $|\hat{\psi}(\omega)|$ the amplitude of $\hat{\psi}(\omega)$. The mother wavelet may be interpreted as the impulse response of a band-pass filter since $\hat{\psi}(0) = \int_{-\infty}^{+\infty} \psi(x)dx = 0$. The mother wavelet is defined jointly with a scaling function (also called father wavelet), who's mean is non-vanishing and which has a

corresponding low-pass filter impulse response, namely $\phi \in L^2(\mathcal{R})$ with

$$\int_{-\infty}^{+\infty} \phi(x)dx = 1. \tag{4.1.3}$$

its Fourier transform also satisfies,

$$|\hat{\phi}(\omega)|^2 = \int_1^{+\infty} |\hat{\psi}(s\omega)|^2 \frac{ds}{s} = \int_\omega^{+\infty} \frac{|\hat{\psi}(\xi)|^2}{\xi} d\xi \tag{4.1.4}$$

By dilating and translating a wavelet function, we obtain a family of wavelet atoms

$$\psi_{u,s}(x) = \frac{1}{\sqrt{s}} \psi \left(\frac{x-u}{s} \right)$$

with $||\psi_{u,s}|| = 1$. The wavelet transform of a function $f \in L^2(\mathcal{R})$ at u and s is defined as

$$Wf(u,s) =< f, \psi_{u,s} >= \int_{-\infty}^{+\infty} f(x) \frac{1}{\sqrt{s}} \psi \left(\frac{x-u}{s} \right) dx = f * \bar{\psi}_s(u) \tag{4.1.5}$$

with

$$\bar{\psi}_s(u) = \psi_s(-u) = \frac{1}{\sqrt{s}} \psi \left(\frac{-u}{s} \right). \tag{4.1.6}$$

$Wf(u,s)$ measures the variation of f in a neighborhood of u, whose size is proportional to s.

Similarly, a family of scaling atoms may be generated as

$$\phi_{u,s}(x) = \frac{1}{\sqrt{s}} \phi \left(\frac{x-u}{s} \right)$$

whose transform at u and s

$$Lf(u,s) =< f, \phi_{u,s} >= \int_{-\infty}^{+\infty} f(x) \frac{1}{\sqrt{s}} \phi \left(\frac{x-u}{s} \right) dx \tag{4.1.7}$$

provides the low-frequency approximation of f at the scale s in a neighborhood of u.

By the low-pass and band-pass filters' properties, the wavelet and scaling atoms play an important role in generating bases of the approximation and detail spaces as shown in the following sections.

4.2 Wavelet frames

The frame theory was originally developed by Duffin and Schaeffer [24] to reconstruct a band-limited signal f from irregularly spaced samples $\{f(t_n)\}_{n\in\mathcal{Z}}$. A frame is a family of vector $\{\phi_n\}_{n\in\Gamma}$ in a space that characterizes any signal f from its inner products $\{< f, \phi_n >\}_{n\in\Gamma}$, where Γ is an index set. The formal definition of a frame is

Definition 3. *(Definition of Frame): The sequence $\{\phi_n\}_{n\in\Gamma}$ is a frame of a Hilbert space H if there exist two constants A and B such that for any $f \in H$*

$$A||f||^2 \leq \sum_{n\in\Gamma} | < f, \phi_n > |^2 \leq B||f||^2 \tag{4.2.1}$$

when $A = B$, the frame is said to be tight. Obviously one can see that an orthonormal basis is a tight frame with $A = B = 1$. These bounds may also display the redundancy of a frame representation of a signal. An operator U defined by $U : Uf[n] =< f, \phi_n >,\ \forall\ n \in \Gamma$ is called a frame operator.

With $\{\phi_n\}_{n\in\Gamma}$ denoting a frame, the sampling observed data represent coefficients, whose reconstruction is carried out using a dual frame $\{\tilde{\phi}_n\}_{n\in\Gamma}$ defined by

$$\tilde{\phi}_n = (U^*U)^{-1}\phi_n, \tag{4.2.2}$$

to yield f as

$$f = \sum_{n\in\Gamma} < f, \phi_n > \tilde{\phi}_n. \tag{4.2.3}$$

If the frame $\{\phi_n\}_{n\in\Gamma}$ is tight, then the dual frame $\{\hat{\phi}_n\}$ is equal to $\{\phi_n\}$ with a scaling difference, which generally simplifies the numerical implementation of a frame decomposition and explains its wider acceptance.

Wavelet frames are constructed by starting with a continuous wavelet transform whose translation and scale parameters are appropriately sampled to cover the time-frequency plane with corresponding discrete wavelet family. To obtain a full cover, we sample the scale parameter s along an exponential sequence $\{a^j\}_{j\in\mathcal{Z}}$, with a sufficiently small dilation step $a > 1$. The time translation u is sampled uniformly at intervals proportional to the scale a^j,

yielding

$$\psi_{j,n}(x) = \frac{1}{\sqrt{a^j}}\psi\left(\frac{x - nu_0a^j}{a^j}\right) \tag{4.2.4}$$

Necessary and sufficient conditions for $\{\psi_{j,n}\}_{(j,n)\in\mathcal{Z}^2}$ to be a frame and have a dual frame is stated (see [21]). However, the sampling interval $a^j u_0$ might cause a translation distortion if its value is large compared to the rate of variations of $f * \bar{\psi}_{a^j}(t)$. To construct a translation invariant wavelet representation, the scale s is discretized while the translation parameter u is not. The scale is sampled along a dyadic sequence $\{2^j\}_{j\in\mathcal{Z}}$. The dyadic wavelet transform of $f \in L^2(\mathcal{R})$ is defined by

$$Wf(u, 2^j) =< f, \psi_{u,2^j} >= \int_{-\infty}^{+\infty} f(x)\frac{1}{\sqrt{2^j}}\psi\left(\frac{x-u}{2^j}\right)dx = f * \bar{\psi}_{2^j}(u) \tag{4.2.5}$$

with

$$\bar{\psi}_{2^j}(u) = \psi_{2^j}(-u) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{-u}{2^j}\right). \tag{4.2.6}$$

It can be shown that the normalized dyadic wavelet transform operator $Uf(j, u) = Wf(u, 2^j)$ satisfies frame inequalities and a reconstructing wavelet $\tilde{\psi}$ may be constructed. In practice, we need to compute a discrete dyadic wavelet transform, which may be carried out by a fast filter bank algorithm for an appropriately designed wavelet and described as follows.

4.3 Wavelet basis

A wavelet dilated by 2^j and translated by $2^j n$ for all $(j, n) \in \mathcal{Z}^{\in}$ generates an orthonormal basis of $L^2(\mathcal{R})$ but distribution information at different resolutions. This is intimately related to multi-resolution signal approximation defined below.

Definition 4. *A sequence $(V_j)_{j\in\mathcal{Z}}$ of closed subspaces of $L^2(\mathcal{R})$ is a multi-resolution approximation if the following 6 properties are satisfied:*

$$\forall\ (j,k)\in\mathcal{Z}^2, f(t)\in V_j \leftrightarrow f(t-2^jk)\in V_j,$$

$$\forall\ j\in\mathcal{Z}, V_{j+1}\subset V_j,$$

$$\forall\ j\in\mathcal{Z}, f(t)\in V_j \leftrightarrow f(t/2)\in V_{j+1},$$

$$\lim_{j\to+\infty} V_j = \cap_{j=-\infty}^{+\infty} V_j = \{0\},$$

$$\lim_{j\to-\infty} V_j = \textit{Closure}\left(\cup_{j=-\infty}^{+\infty} V_j\right) = L^2(\mathcal{R}).$$

There exists θ such that $\theta(t-n)_{n\in\mathcal{Z}}$ is a Riesz basis [1] of V_0.

We need to mention that V_j characterizes the signal approximation at the resolution 2^{-j}, The approximation of f on V_j is defined as the orthogonal projection $P_{V_j}f$ of function f onto the space V_j.

In order for a family $\{\theta(t-n)\}_{n\in\mathcal{Z}}$ to be a Riesz basis of the space V_0, the necessary and sufficient condition is there exist $A>0,\ B>0$, such that

$$\frac{1}{B}\le\sum_{k=-\infty}^{+\infty}|\hat{\theta}(\omega-2k\pi)|^2\le\frac{1}{A}. \tag{4.3.1}$$

This yields the construction of a scaling function ϕ through Fourier transform of $\theta(t)$

$$\hat{\phi}(\omega)=\frac{\hat{\theta}(\omega)}{(\sum_{k=-\infty}^{+\infty}|\hat{\theta}(\omega-2k\pi)|^2)^{1/2}}, \tag{4.3.2}$$

for which we denote

$$\phi_{j,n}(t)=\frac{1}{\sqrt{2^j}}\phi\left(\frac{t-2^jn}{2^j}\right). \tag{4.3.3}$$

Scaling atoms $\{\phi_{j,n}\}_{n\in\mathcal{Z}}$ defined above normally span a multi-resolution approximation space V_j. As we know that $V_j\subset V_{j-1}$, let W_j be the orthogonal complement of V_j in V_{j-1}:

$$V_{j-1}=V_j\oplus W_j$$

the orthogonal projection of f on V_{j-1} can be further decomposed as

$$P_{V_{j-1}}f=P_{V_j}f\oplus P_{W_j}f.$$

[1] Definition of Riesz basis can be found in [66].

The complement $P_{W_j}f$ provides further "details" of f at scale 2^j and one can construct an orthonormal basis of W_j by wavelet atoms $\{\psi_{j,n}\}_{n\in\mathcal{Z}}$, which are obtained by sampling u and s on a 2^j grid,

$$\psi_{j,n}(x) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{x-2^j n}{2^j}\right). \tag{4.3.4}$$

With $\{\phi_{j,n}\}_{n\in\mathcal{Z}}$ and $\{\psi_{j,n}\}_{n\in\mathcal{Z}}$ as the orthonormal bases of V_j and W_j, the coefficients of f in these spaces are given by

$$a_j[n] =< f, \phi_{j,n} > \quad \text{and} \quad d_j[n] =< f, \psi_{j,n} > \tag{4.3.5}$$

The orthogonal projection of f over V_j and W_j are therefore obtained in the following expressions,

$$\begin{aligned} P_{V_j}f &= \sum_{n=-\infty}^{\infty} < f, \phi_{j,n} > \phi_{j,n} = \sum_{n=-\infty}^{\infty} a_j[n]\phi_{j,n}, \\ P_{W_j}f &= \sum_{n=-\infty}^{\infty} < f, \psi_{j,n} > \psi_{j,n} = \sum_{n=-\infty}^{\infty} d_j[n]\psi_{j,n}. \end{aligned} \tag{4.3.6}$$

A multi-resolution approximation $\{V_j\}_{j\in\mathcal{Z}}$ is entirely characterized by the scaling function ϕ that generates an orthonormal basis for each space V_j, This may in turn be shown to be implemented by a conjugate mirror filter.

4.3.1 Wavelet design: Connection to conjugate mirror filters

Let h and g be a pair of finite impulse response filters, h is a low-pass filter whose transfer function satisfies $\hat{h}(0) = \sqrt{2}$, where $\hat{h}(\omega) = \sum_{n=-\infty}^{+\infty} h[n]e^{-in\omega}$ is the Fourier series of the discrete filter $h[n]$.

Mallat[64] and Meyer[69] point out that, for an integrable scaling function, the Fourier series of $h[n]$ satisfies

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega+\pi)|^2 = 2, \quad \forall\omega \in \mathcal{R} \tag{4.3.7}$$

and

$$\hat{h}(0) = \sqrt{2}. \tag{4.3.8}$$

A discrete filter whose Fourier series satisfies Eq.(4.3.7) is called a conjugate mirror filter.

Conversely, they also give necessary conditions on which a scaling function can be constructed, namely, if $\hat{h}(\omega)$ is 2π periodic and continuously differentiable in a neighborhood of $\omega = 0$, and it satisfies Eq.(4.3.7) and Eq.(4.3.8), and if

$$\inf_{\omega\in[-\pi/2,\pi/2]} |\hat{h}(\omega)| > 0, \tag{4.3.9}$$

then

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} = \frac{1}{\sqrt{2}}\hat{h}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2}) \tag{4.3.10}$$

is the Fourier transform of a scaling function $\phi \in L^2(\mathcal{R})$. Further, we can decompose

$$\frac{1}{\sqrt{2}}\phi(t/2) = \sum_{n=-\infty}^{+\infty} h[n]\phi(t-n) \tag{4.3.11}$$

with

$$h[n] =< \frac{1}{\sqrt{2}}\phi(\frac{t}{2}), \phi(t-n) > . \tag{4.3.12}$$

Suppose that the Fourier transform of ϕ is finite, the corresponding wavelet ψ is defined through a Fourier Transform as

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}}\hat{g}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2}), \tag{4.3.13}$$

with

$$\hat{g}(\omega) = e^{-i\omega}\hat{h}^*(\omega+\pi). \tag{4.3.14}$$

The necessary and sufficient conditions on $\hat{g}$ such that, for any scale 2^j, $\{\psi_{j,n}(x)\}_{n\in\mathcal{Z}}$ be an orthonormal basis of W_j and $\{\psi_{j,n}\}_{(j,n)\in\mathcal{Z}^2}$ be an orthonormal basis of $L^2(\mathcal{R})$ is

$$|\hat{g}(\omega)|^2 + |\hat{g}(\omega+\pi)|^2 = 2 \tag{4.3.15}$$

and

$$\hat{g}(\omega)\hat{h}^*(\omega) + \hat{g}(\omega+\pi)\hat{h}^*(\omega+\pi) = 0. \tag{4.3.16}$$

It is shown that $\hat{g}(\omega)$ is the Fourier series of

$$g[n] =< \frac{1}{\sqrt{2}}\psi(\frac{t}{2}),\ \phi(t-n) >, \tag{4.3.17}$$

which are the decomposition coefficients of

$$\frac{1}{\sqrt{2}}\psi(\frac{t}{2}) = \sum_{n=-\infty}^{+\infty} g[n]\phi(t-n), \tag{4.3.18}$$

for which, calculating the inverse Fourier transform of Eq.(4.3.14) yields

$$g[n] = (-1)^{1-n}h[1-n] \tag{4.3.19}$$

Therefore g is a high-pass conjugate mirror filter. The conjugate mirror filters $h[n]$, $g[n]$ play an important role in the fast wavelet transform algorithm. A fast filter bank algorithm is further introduced to calculate the coefficients of a signal measured at a finite resolution, thus, the following formula is further proceeded to Eq. (4.3.5)

$$a_{j+1}[p] = \sum_{n=-\infty}^{+\infty} h[n-2p]a_j[n]$$

$$d_{j+1}[p] = \sum_{n=-\infty}^{+\infty} g[n-2p]a_j[n].$$

We can see that signals are decomposed into low-pass and high-pass components sub-sampled by 2, and coefficient $a_j[p]$ may be reconstructed as

$$\begin{aligned} a_j[p] &= \sum_{n=-\infty}^{+\infty} h[p-2n]a_{j+1}[n] + \sum_{n=-\infty}^{+\infty} g[p-2n]d_{j+1}[n] \\ &= \check{a}_{j+1} \star h[p] + \check{d}_{j+1} \star g[p] \end{aligned} \tag{4.3.20}$$

where " $\star$ " represents convolution and

$$\check{x}[n] = \begin{cases} x[k] & \text{if } n = 2k \\ 0 & n = 2k+1 \end{cases}$$

4.3.2 Vanishing Moments vs. Support size of a wavelet

In order to parsimoniously represent a function f in a wavelet basis, fewer non-zero coefficients are desirable. This is depend on the regularity of the function f, and on choosing an appropriate wavelet, namely, wavelet with specific vanishing moments and support size.

A wavelet with n vanishing moments satisfies the following formula,

$$\int_{-\infty}^{+\infty} x^k \psi(x) dx = 0,\ 0 \leq k < n, \tag{4.3.21}$$

which means that a wavelet ψ is orthogonal to any polynomial of degree up to $n-1$. This also leads to the following relation to its Fourier transform $\hat{\psi}$,

$$\int_{-\infty}^{+\infty} x^k \psi(x) dx = i^k \hat{\psi}^{(k)}(0) = 0,\ 0 \leq k < n \tag{4.3.22}$$

where $i = \sqrt{-1}$. This property reveals that, if a wavelet ψ has n vanishing moments, $\hat{\psi}$ and its first $n-1$ derivatives are 0 at $\omega = 0$. Therefore, if $f \in C^n$, C^n is a collection of functions which are n times continuously differentiable (i.e. f can be well approximated by a Taylor polynomial), then such an analyzing wavelet produces small amplitude coefficients at fine scales and one can shown that the decay rate of $|Wf(u,s)|$ across scales s is closely related to the exponential degree of the Lipschitz regularity of f. This property is used to detect singularities by finding abscissa where modulus maxima converge at fine scales, as discussed in detail in [69, 45, 67].

One may also argue that the coefficients $< f, \psi_{j,n} >$ may have large amplitudes if a wavelet ψ has a large support, To therefore minimize the number of high amplitude coefficients, we need to reduce the support size of ψ as much as possible. The constraints imposed on orthogonal wavelets imply that if ψ has p vanishing moments then its support is at least of size $2p-1$ when choosing a particular wavelet [33, 88]. We hence face a tradeoff between the number of vanishing moments and the support size.

4.4 Daubechies Wavelets

Daubechies wavelets have minimum size supports for any given vanishing moments of order p. It is shown that the support sizes of a scaling function ϕ and a wavelet ψ are related to the conjugate mirror filter h that are used to construct them. The scaling function ψ has a compact support if and only if h has a compact support, and furthermore their support are equal. If the support of h and ϕ is $[N_1, N_2]$ then the support of ψ is $[(N_1 - N_2 + 1)/2, (N_2 - N_1 + 1)/2]$(see [64]). In addition, to ensure that a wavelet has p vanishing moments, the

Fourier transform $\hat{h}$ of the conjugate mirror filter h, which is used to construct ψ, must have a zero of order p at $\omega = \pi$, namely, $\hat{h}(\omega)$ can be written as

$$\hat{h}(\omega) = \sqrt{2}\left(\frac{1+e^{-i\omega}}{2}\right)^p R(e^{-i\omega}), \tag{4.4.1}$$

where $R(e^{-i\omega})$ is a polynomial of minimum degree m such that $\hat{h}$ satisfies

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega+\pi)|^2 = 2. \tag{4.4.2}$$

As a result, h has $N = m + p + 1$ non-zero coefficients. Thus obtaining a minimum support wavelet is equivalent to obtaining a minimum support h, and the minimum degree m of R required is $m = p - 1$, see Daubechies[20].

Daubechies wavelets are constructed by choosing a minimum degree polynomial

$$R(e^{-i\omega}) = \sum_{k=0}^{m} r_k e^{-ik\omega} = r_0 \prod_{k=0}^{m}(1 - a_k e^{-i\omega}) \tag{4.4.3}$$

such that $|R(e^{-i\omega})|^2 = P(\sin^2(\omega/2))$, where $P(x)$ is a polynomial satisfying

$$(1-x)^p P(x) + x^p P(1-x) = 1 \tag{4.4.4}$$

It turns out that Daubechies wavelets constructed above have a minimum size support equal to $[-p+1, p]$. Daubechies wavelets are, however, very asymmetric, as shown in Fig.(4.1).

Daubechies wavelets demonstrate that there is a tradeoff between vanishing moments and the support size, specifically the higher the vanishing moments, the larger the support size. The shortest support Daubechies wavelet is a Haar wavelet, which has a vanishing moment of order 1.

4.5 Wavelet Packet

Instead of decomposing only the approximation spaces V_j to construct lower resolution approximation space V_{j+1} and detail space W_{j+1} by wavelet bases, we can also decompose the detail space W_j into two subspaces, an approximation and a detail space. This calls for

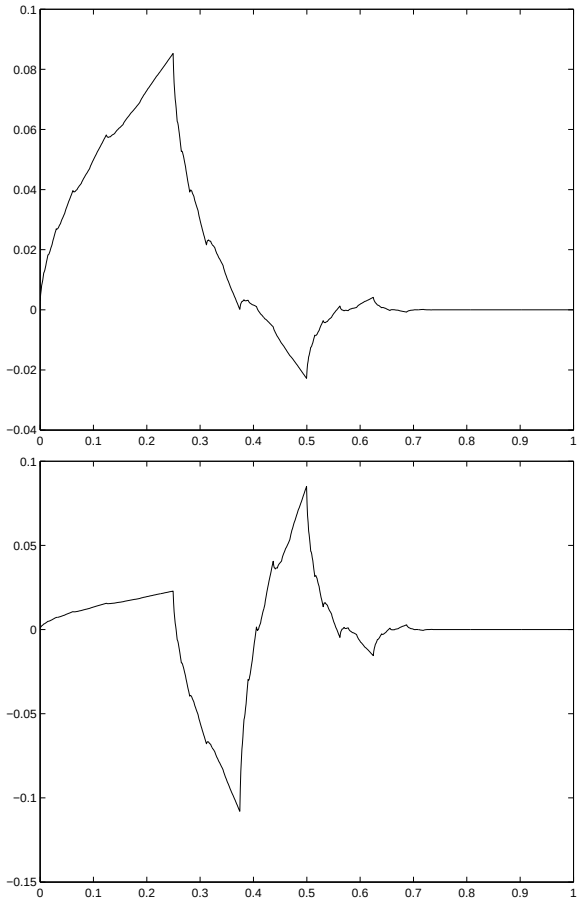

Figure 4.1: Daubechies scaling function ϕ (top) and wavelet ψ(bottom) with vanishing moments 2

new bases, the so-called wavelet packets, due to Coifman, Meyer and Wickerhauser[79]. The analysis may be interpreted as spectral partitioning.

If the signals are approximated at scale 2^L, we can represent the recursive splitting of vector spaces in a binary tree. The root of the tree is associated to the approximation space V_L. Denote $W_L^0 = V_L$, each node of the tree is labelled by $(j,\ p)$, and is associated to a detail space W_j^p, where $j - L \geq 0$ is the depth of the node in the tree , and p is the number of nodes that are on its left at the same depth $j - L$. Therefore, the two children nodes of $(j,\ p)$ are orthogonal subspaces (Wavelet packet spaces) such that

$$W_j^p = W_{j+1}^{2p} \oplus W_{j+1}^{2p+1}$$

where W_{j+1}^{2p} and W_{j+1}^{2p+1} may be viewed as the approximation and detail spaces of W_j^p and the two wavelet packet orthogonal bases of the two children notes are defined as

$$\psi_{j+1}^{2p}(t) = \sum_{n=-\infty}^{+\infty} h[n]\psi_j^p(t - 2^j n) \tag{4.5.1}$$

and

$$\psi_{j+1}^{2p+1}(t) = \sum_{n=-\infty}^{+\infty} g[n]\psi_j^p(t - 2^j n) \tag{4.5.2}$$

with

$$h[n] =< \psi_{j+1}^{2p}(u),\ \psi_j^p(u - 2^j n) >, \quad g[n] =< \psi_{j+1}^{2p+1}(u),\ \psi_j^p(u - 2^j n) > \tag{4.5.3}$$

This recursive splitting defines a binary wavelet packet tree where each parent node is divided into two orthogonal subspaces each with concentrated energy in different frequency bins. The idea of wavelet packets generalizes the link between multiresolution approximations and wavelets, and the designed bases for wavelet packets are well adapted to decomposing signals that have different behavior in different frequency intervals, several specific wavelet packet bases, such as cosine bases, block bases, lapped orthogonal bases, etc. are intensively discussed, see[64, 92, 80, 79, 68, 65].

Chapter 5

Wavelet Frame-Based Nonlinear Filtering

5.1 Introduction

In spite of the fact that the performance improvement as in Chapter 3 was remarkable, the drawback was the loss of features such as texture which, in some class(other than those investigated in [89]) of images is very important to preserve.

the nature of question which then follows and which is addressed in this chapter is whether an efficient and effective filtering approach can be made feature (e.g. texture)preserves, here we address this problem and show that using wavelet frames with wavelets of higher order moments than Haar's is tantamount to accounting for longer term correlation structure while preserving the local focus. This hence yields an efficient tool in analyzing and enhancing images with a careful account for texture information.

we explain a connection between the equation and the process to Haar wavelet coefficient to establish a direct equivalence between a linear Heat equation and a Haar coefficients based evolution.

In light of this derivation shown in Eq. 5.4.1, we proceed to generalize this connection and in fact derive wavelet frame coefficients, this leads to a remarkable signal and image

enhancement while preserved features which are important to other approaches, such as, image classification etc.

5.2 Problem Statement

As noted above the Perona-Malik equation still enjoys a great deal of popularity for achieving a selective nonlinear filtering, compatible with the desired objective of image filtering, namely that homogeneous areas be maximally smoothed while edge contours be maximally preserved (or equivalently minimally smoothed). It is expressed as

$$\frac{\partial U(t,x)}{\partial t} = div\left(\mathcal{F}\left(\mid \nabla\left(U(t,x)\right)\mid\right)\nabla U(t,x)\right), \tag{5.2.1}$$

where $\mathcal{F}(v)$ may be chosen as $\mathcal{F}(v) = e^{-\frac{v^2}{K^2}}$, K determines the rate of decay and thus the extent of smoothing of $U(t,x)$ for a given gradient size. Many other techniques have been proposed with each addressing different aspects of the limitations of the above equation. Specifically, one which addresses the stopping criterion problem [50, 49] may be written for simplicity in a 1-D evolution as a Markov chain equation, see Eq.(3.4.2).

To the best of our knowledge, none of these techniques [89, 50] resolves the problem of texture loss alluded to in the introduction. The fact that a first order difference implementation of a gradient in the selective filtering of an image is a main source of this loss may easily be observed by the convergence of the data to staircase functions[49] and can be seen in Fig. 5.1. Our goal in the sequel is in effect to lift this limitation by reinterpreting the first difference implementation as a Haar wavelet coefficient and by subsequently seeking a more regular wavelet implementation/approximation as we elaborate further next.

5.3 A Multiscale Approach to Scale-Space Analysis

Much of the research in nonlinear diffusion [89] has been carried out for the most part on a parallel track to and with little interaction with all that had been pursued in wavelet or multiresolution analysis. This is in spite of the fact that both approaches are very much based on the notion of scale, and that both fully use information gleaned along it. Creative

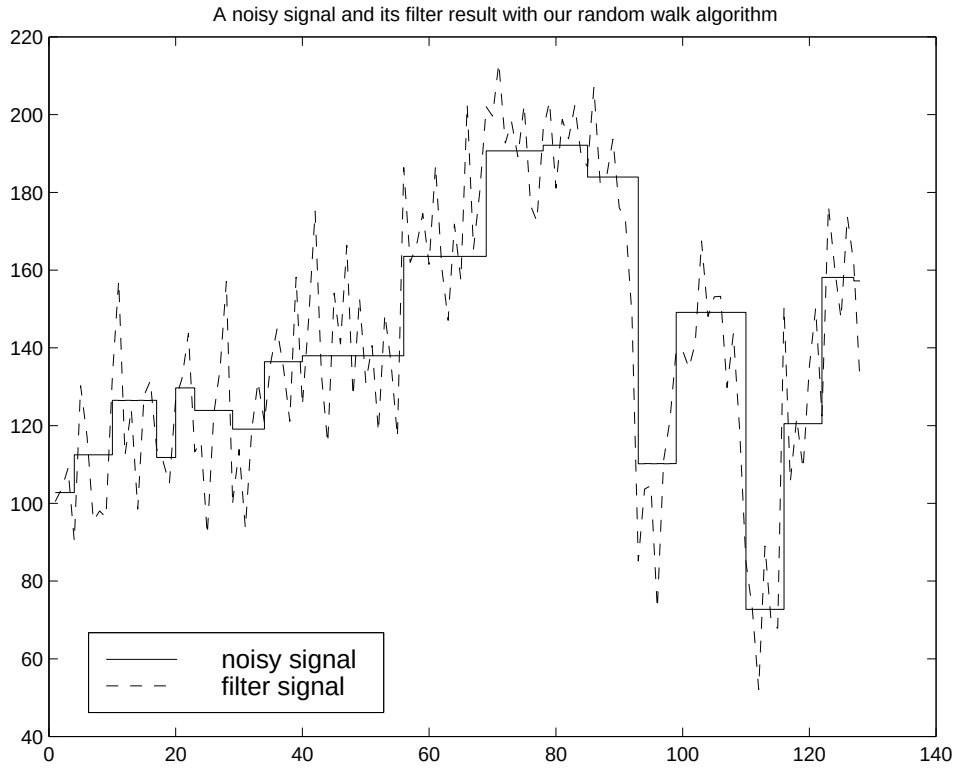

Figure 5.1: A profile of noisy Lenna image and filtered result with Random walk algorithm[52].

and clever ideas germane to the two philosophies resulted, and distinct advantages and limitations emerged. A natural question which then arises is on their interplay and on any potential gain which may result if the synergy is exploited. Towards that end, we first recall some facts about frames and their role in signal representations and subsequent filtrations.

5.4 Frame Representation and Reconstruction

As is well known, an image may be well represented in a Haar wavelet frame by obviating the dyadic down-sampling step (i.e. a redundant representation) of the usual orthonormal representation. While the latter representation is usually parsimonious and yields a perfect reconstruction, it exhibits a visually noticeable loss of information in the course of additional transformations such as coefficient thresholding for denoising or compression, etc. This loss is particularly evident in an orthonormal representation with a short support analysis

wavelet, and may be attributed to the fact that any given transformed coefficient has a significant local influence and hence impact on the visual outcome . Smoothing by coefficient thresholding which is usually popular in wavelet-based denoising [52, 50], is in contrast to the softer and more progressive smoothing commonly encountered in nonlinear diffusion of scale space filtering [75]. The redundancy of this continuous scale approach in some sense, counterbalances the singular effect of an orthonormal wavelet coefficient.

Our goal in this section is to establish for a given signal an equivalence between smoothing based on the Heat equation and that based on a diffusion-emulating transformation of its frame coefficients. The wealth of available wavelet functions makes it possible for us to optimize our desired ability to capture longer term (higher than Markov of order 1) correlation information at a given spatial location in an image and to still preserve the local focus critical to exploiting salient features in nonlinear filtering.

To proceed, let $\phi(x)$ be a scaling function with a compact support such that$\{\phi(x-n)\}_{n\in\mathbb{Z}}$ is an orthonormal basis of V_0, the space of observations. Its Fourier transform is $\hat{\phi}(\omega) = \frac{1}{\sqrt{2}}\hat{h}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2})$, and consequently satisfies the multiresolution analysis framework [66]. Denote by $\psi(x)$ a function with a Fourier transform $\hat{\psi}(\omega) = \frac{1}{\sqrt{2}}\hat{g}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2})$ i.e., a corresponding and so-called mother wavelet. As is well known [66], we can write

$$\frac{1}{\sqrt{2}}\phi(\tfrac{x}{2}) = \sum_{-\infty}^{+\infty} h(n)\phi(x-n); \quad \frac{1}{\sqrt{2}}\psi(\tfrac{x}{2}) = \sum_{-\infty}^{+\infty} g(n)\phi(x-n),$$

where $\{h(n)\}_{n\in\mathbb{N}}$ and $\{g(n)\}_{n\in\mathbb{N}}$ with Fourier transforms $\hat{h}(\omega), \hat{g}(\omega)$ satisfy the complementarity property $\hat{g}(\omega) = e^{-i\omega}\hat{h}^*(\omega+\pi)$.

Define

$$\begin{cases} \phi_{j,n}(x) = \frac{1}{\sqrt{2^j}}\phi\left(\frac{x-2^{j-1}n}{2^j}\right) \\ \psi_{j,n}(x) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{x-2^{j-1}n}{2^j}\right) \end{cases}$$

with $\{\phi_{j,n}(x)\}$ and $\{\psi_{j,n}(x)\}$ as tight frames of V_j and W_j (the so-called approximation and detail subspaces). It follows that $\{\phi_{j,2n}(x)\}$ and $\{\phi_{j,2n+1}(x)\}$ as well as $\{\psi_{j,2n}(x)\}$ and $\{\psi_{j,2n+1}(x)\}$ are respectively orthonormal bases of V_j and W_j. With the selected functions in hand, and upon obtaining a frame representation of a signal, we first proceed to show

how it may be used to implement a linear filter equivalent to that achieved by a linear heat equation.

Proposition 3. *The numerical implementation of a diffusion effect of Laplacian $\Delta(\cdot)$ operating on a signal $U(x)$ (as in the linear heat equation), may be achieved by an iterative subtraction of the highest detail (also referred to as detail of detail) contribution of a Haar wavelet packet frame representation of the signal.*

$$\begin{aligned} U(n+1,x) &= U(n,x) - (-\frac{1}{2}U^{dd}(n,x-1)) \\ &= U(n,x) + \frac{1}{2}U^{dd}(n,x-1) \end{aligned} \tag{5.4.1}$$

■

Prior to proving this proposition, we have to establish the following two lemma.

Lemma 1. *If we are given a function $f(x)$ together with its frame representation coefficients,*

$$a_j(n) =< f, \phi_{j,n} > \text{ and } d_j(n) =< f, \psi_{j,n} >$$

in a tight frame $\{\phi_{j,n}(x)\}, \{\psi_{j,n}(x)\}, (n,j) \in \mathbb{Z}$ the following formulae yield

$$\begin{aligned} a_{j+1}(p) &= \sum_{-\infty}^{+\infty} h(n-p)a_j(n) \text{ and} \\ d_{j+1}(p) &= \sum_{-\infty}^{+\infty} g(n-p)a_j(n), \end{aligned} \tag{5.4.2}$$

with either of the following two reconstructions for $a_j(p)$

$$a_j(p) = \sum_{-\infty}^{+\infty} a_{j+1}(2n)h(p-2n) + \sum_{-\infty}^{+\infty} d_{j+1}(2n)g(p-2n) \tag{5.4.3}$$

or

$$a_j(p) = \sum_{-\infty}^{+\infty} a_{j+1}(2n+1)h(p-2n-1) + \sum_{-\infty}^{+\infty} d_{j+1}(2n+1)g(p-2n-1) \tag{5.4.4}$$

Proof: The proof is immediate by noting that

$$\begin{aligned} & < \phi_{j,p},\ \phi_{j,2n} > \\ = & \int \frac{1}{\sqrt{2^{j+1}}}\phi\left(\frac{x-2^j p}{2^{j+1}}\right)\frac{1}{\sqrt{2^j}}\phi\left(\frac{x-2^j n}{2^j}\right)dx \\ = & \int \frac{1}{\sqrt{2}}\phi\left(\frac{u-p}{2}\right)\phi(u-n)du \\ = & \ h(n-p), \end{aligned} \tag{5.4.5}$$

and similarly $< \psi_{j+1,p},\ \phi_{j,2n} >= g(n-p)$. ∎

Recall that our goal is to establish a direct relationship between a linear diffusion filter and its numerical implementation in a wavelet domain. The space/time invariance property of the linear diffusion imposes a frame-based representation of a signal being analyzed. The choice of a wavelet frame representation for this purpose is further justified by the intrinsic analytic property of wavelets for focusing useful energy (e.g. of the desired signal) in relatively few coefficients and for spreading that of the noise over many coefficients. This consequently indicates that an efficient and systematic multiscale representation with a sufficient and flexible spectral redundancy may result. Towards that end, we start by stating the following,

Lemma 2. *The detail space W_j can be expressed as a direct sum of two subspaces $W_j = V_{j,L} \bigoplus W_{j,L}$. Defining*

$$\begin{aligned} \psi^a_{j,p}(x) = & \sum_{-\infty}^{+\infty} h(n-p)\psi_{j,n}(x) \ \textit{and} \\ \psi^d_{j,p}(x) = & \sum_{-\infty}^{+\infty} g(n-p)\psi_{j,n}(x), \end{aligned} \tag{5.4.6}$$

$\{\psi^a_{j,p}(x)\}$ and $\{\psi^d_{j,p}(x)\}$ are respectively frames of $V_{j,L}, W_{j,L}$ while $(\{\psi^a_{j,2n}(x)\}, \{\psi^d_{j,2n}(x)\})$ and $(\{\psi^a_{j,2n+1}(x)\}, \{\psi^d_{j,2n+1}(x)\})$ are respectively the corresponding orthonormal bases of $V_{j,L}$ and $W_{j,L}$. Furthermore denoting the coefficients of the decomposition of the details in the frames $\{\psi^a_{j,p}(x)\}$ and $\{\psi^d_{j,p}(x)\}$ by $\{da_j(n)\}$ and $\{dd_j(n)\}$, we have per

Eq.(5.4.4), the following reconstruction relationship

$$d_j(n) = \sum_{-\infty}^{+\infty} da_j(2n)h(p-2n) + \sum_{-\infty}^{+\infty} dd_j(2n)g(p-2n)$$

or

$$d_j(n) = \sum_{-\infty}^{+\infty} da_j(2n+1)h(p-2n-1) + \sum_{-\infty}^{+\infty} dd_j(2n+1)g(p-2n-1).$$

Proof: Similar to the proof of Lemma 1.

Proof of Proposition 1: See Appendix A. ∎

5.5 Selection and Impact of a Wavelet Support

The Fourier transform $\hat{U}_1^d(n,\omega)$ may be written in a Haar basis as

$$\hat{U}_1^d(n,\omega) = \hat{g}^*(\omega)\hat{U}(n,\omega),$$

while that of Eq. (5.4.1) may be written as,

$$\hat{U}(n+1,\omega) = (1+\frac{1}{2}\hat{g}^*(\omega)\hat{g}^*(\omega)e^{-j\omega})\hat{U}(n,\omega) \triangleq LF(\omega)\hat{U}(n,\omega) \tag{5.5.1}$$

where $LF(\cdot)$ is a low pass filter resulting from Eq. (5.4.1). The transfer function corresponding to a Haar wavelet $\hat{g}(\omega)$ with its conjugate denoted by $\hat{g}^*(\omega)$ is given by

$$g(\omega) = \sqrt{2}sin(\frac{\omega}{2})e^{j\frac{\pi-\omega}{2}},$$

hence resulting in

$$LF(\omega) = (1+cos\omega)/2.$$

As may be suggested by Eqs. (5.4.1,5.5.1) as well our earlier comments, we may select a different $g(\cdot)$ (e.g. a higher order wavelet, such as a Daubchies-4 etc.) and investigate the overall behavior of $LF(\omega)$ as illustrated in Fig. 5.2.

The choice the number of vanishing moments is a degree of freedom which may be optimized around specific applications or goals. In Fig. 5.2, we contrast the characteristics of our diffusion filter using a Haar-based implementation of the Laplacian operator to that based on a higher order wavelet such as Daubichies wavelet with vanishing moments $4(D4)$. The

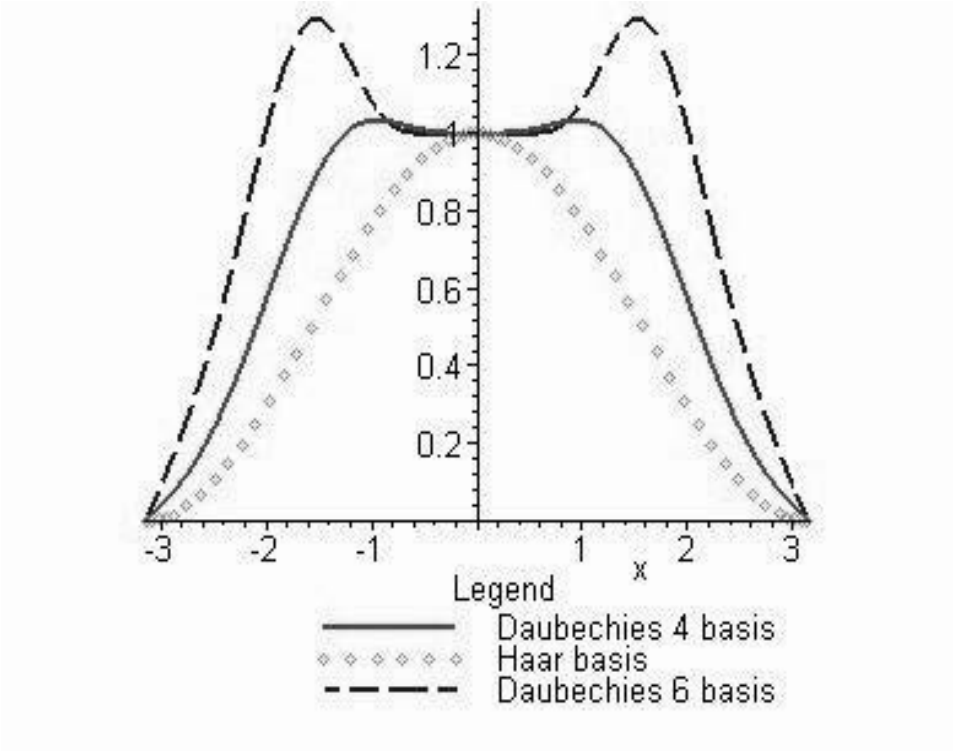

Figure 5.2: Spectral characteristics of Heat equation-like filter using Haar and Daubechies-4/6 wavelet functions.

larger support wavelet ($D4$) exhibits a more graceful but nevertheless sharp roll-off of the lowpass filter. This is a direct result of the selected support and hence of the smoothness of the wavelet. The selection of higher order wavelets may be justified as a solution to the blockyness problem pointed out in Section I on several counts. Recall that our frame-based diffusion amounted to a progressive reduction/elimination of specific frame coefficients (high details). The linearity of this filtering effectively assumes that the wavelet (i.e. orthogonal) coefficients are uncorrelated which as well known, is inaccurate and leads to either a significant feature distortion or other undesired artifacts in an image reconstruction. This thus highlights the importance of either accounting for the inter-coefficient correlation and avoiding undesired leakage of useful information in incorrectly deleted coefficients or of decorrelating the orthogonal coefficients which may be achieved to a large extent by optimizing the wavelet support. Increasing an analyzing wavelet support (i.e., adopting a higher order wavelet $\psi_i(x)$)), has been shown by Tewfik and Kim[90] to yield increased decorrelation among the wavelet (orthogonal) coefficients. This in turn, enhances the performance of techniques such as wavelet thresholding [23, 51, 84] and diffusion. In concert with the wavelet support selection, the redundancy of a frame operator which yields a range rank reduction, affords additional noise elimination while preserving useful features, such as image texture, as is the goal herein.

An alternative and perhaps more intuitively appealing justification of using higher order wavelets than Haar (for improved filtered reconstruction) follows upon recalling the assumed observed model $f(x) = s(x) + n(x)$ which, for simplicity is assumed to be 1-D.

Fact: *Increasing the support of an analyzing wavelet in the above frame-based diffusion, slows down the smoothing of smooth/polynomial trends.*

The underlying signal of interest $s(x)$ is an a.s. continuous function with discontinuities, which may always be represented as

$$s(x) = \sum_i s_i^c(x) I_{A_i} + \sum_i s^d(x)_i I_{A_i},$$

where $s_c(x) = \sum_i s_i^c(x) I_{A_i}$ and $\sum_i s^d(x)_i I_{A_i}$ respectively are the continuous and the discontinuous parts in interval A_i, such that $\mathcal{I} = \bigcup_i A_i$, and I_{A_i} is an indicator function on the interval. It is well known that $s_i^c(x)$ may be arbitrarily well approximated by a polynomial $p_i(x)$(see Fig.(5.3) for an illustrative example with a continuous function over the partition

of interval and isolated discontinuities) such that

$$\mid s_i^c(x) - p_i(x) \mid \leq \epsilon \quad \text{for a small enough } \epsilon \ .$$

On any interval A_i and $\forall\ x \in A_i$, we have

$$\begin{aligned} f(x) &= s_i^c(x) + s_i^d(x) + n(x) \\ &= p_i(x) + s_i^c(x) - p_i(x) + s_i^d(x) + n(x) \\ &= p_i^k(x) + p_i^{kc}(x) + s_i^c(x) - p_i(x) + s_i^d(x) + n(x) \\ &\stackrel{\Delta}{=} p_i^k(x) + w_i(x), \end{aligned} \tag{5.5.2}$$

where $p_i^k(x)$ is a polynomial of order k and $p_i^{kc}(x) = p_i(x) - p_i^k(x)$, $k \in \mathbb{N}$, and $\| \widetilde{w_i(x)} \| \leq \| p_i^{kc}(x) + \epsilon + s_i^d + n(x) \|$, where $\| \cdot \|$ is the L^2 norm. Given an analysis wavelet $\widetilde{\psi_j^k(x)}$ with k vanishing moments and minimum support, its application to the signal $f(x)$, results in the following coefficients

$$\begin{aligned} d_{i,j}^f &= \int_{A_i} f(x)\psi(x-j)dx = d_j^{p_i^k} + d_j^{w_i}, and \\ d_j^{p_i^k} &= 0, \end{aligned} \tag{5.5.3}$$

where "i" denotes the analysis interval A_i and "j" the translation parameter. Recall that the progressive filtering of $f(x)$ by the frame-based linear diffusion entailed the smoothing of the details of $\sum d_{i,j}^f \widetilde{\psi_j^k(x)} = f_i^d(x) = \sum d_{i,j}^w \widetilde{\psi_j^k(x)}$,

$$f_{ik}^{dd}(x) = \sum_j < f_i^d(x), \widetilde{\psi_j^k(x)} > \widetilde{\psi_j^k(x)} = \sum_j dd_{i,j}^f \widetilde{\psi_j^k(x)} = w_i^{dd}(x)$$

Recall that most of the noise energy in $f_{ik}(x), \forall\ i$, is projected onto the subspace which includes $f^{dd}(x)$ and the latter gets systematically smoothed away from $f(x)$ according to Eq. **??**eq:basicdiffusion), $\forall\ x \in A_i$

$$f_k^o(x) = f(x) - f_{ik}^{dd}(x) = p_i^k(x) + w_i^a(x), \tag{5.5.4}$$

which when iterated leads to

$$f_k^n(x) = f_k^{n-1}(x) - f_k^{(n-1)dd}(x) = p_i^k(x) + w_{in}^a(x), \tag{5.5.5}$$

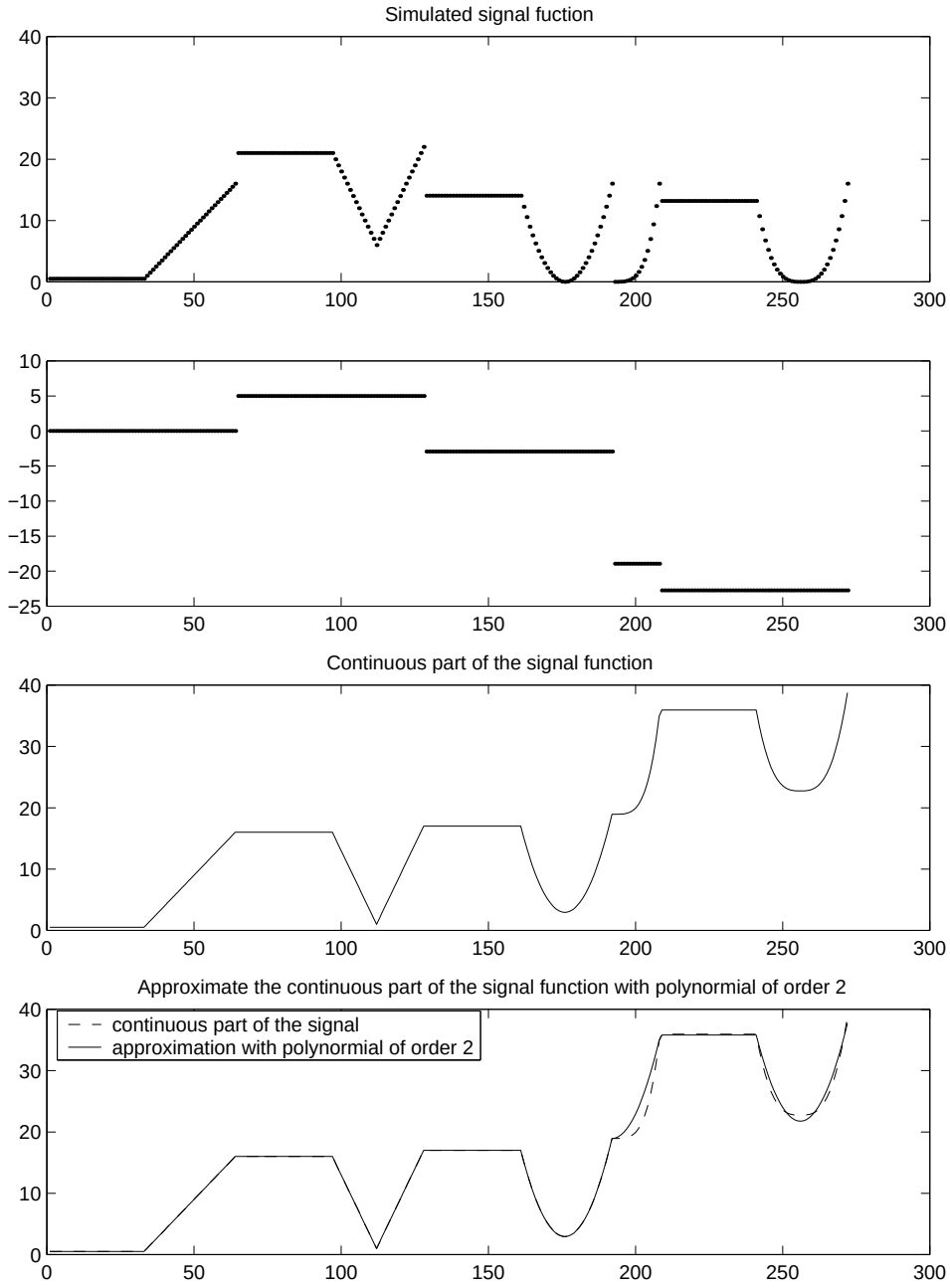

Figure 5.3: A discontinuous signal and its decomposition as sum of continuous part and discontinuous part, also approximation of the continuous part.

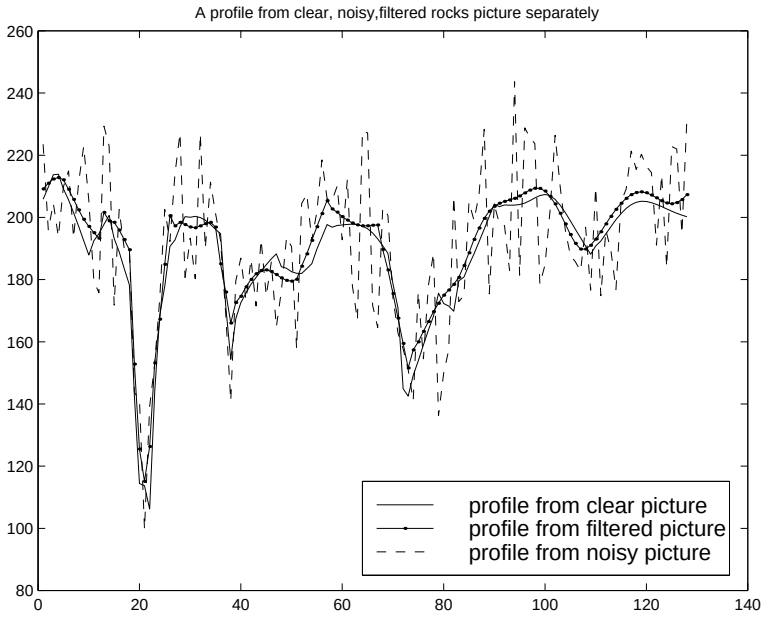

Figure 5.4: A profile take from the rock texture image, with filtered result

$\forall\ x \in A_i$ at step n, where w_i^{ak} represents the iteratively updated information of $w_i(x)$ at step n, or $w_{in}^{a} = w_i(x) - w_{in}^{dd}(x)$, amounting to saying that w_{in}^{dd} (details of details), approach 0 as $n \to \infty$. The evaluation of estimation error may be given as

$$\epsilon^k = \| f^n(x) - s(x) \| \tag{5.5.6}$$

which is seen to decrease as $k' > k$, as $\| s(x) - p^k(x) \| > \| s(x) - p^{k'}(x) \|$, where $p^k(x)$ is the integrated contributions of all polynomials of order k over all intervals in the partition. This hence clearly shows that if $d_j^{p_i^k} = 0, i = 1, \cdots N$, as would be the case for a proper choice of vanishing moments of the analyzing wavelet, the continuous signal contribution to $w_i(x)$ would vanish and equivalently the preservation of all the continuous trends (polynomials) are projected onto the approximation subspace as demonstrated in Fig.(5.4).

5.6 Image Reconstruction using a Haar Frame

To further investigate the interplay between PDE-based filtering and multiscale analysis, we proceed to specialize the foregoing development to a Haar wavelet frame and subsequently derive an equivalent diffusion transformation similar to that of a Heat equation. For clarity

of notation as well algebraic expediency, we adopt a matrix formalism which is convenient for and compatible with an image representation as a matrix. It is also readily extended to any wavelet function which may be selected for the application at hand. It is well known that a nonorthogonal Haar representation of a signal may still yield a reconstruction. To demonstrate such a procedure, denote the impulse response of filters corresponding to a Haar wavelet analysis by $h = [h(0), h(1)] = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$, and $g = [g(0), g(1)] = [-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$. We next construct a $N \times N$ circulant matrix from a vector $[a(1), a(2), \cdots, a(m), 0, \cdots, 0]_{1\times N}$ as $Cir[a(1), \cdots, a(m)]_{N\times N}$, and also write $I_{k,N}$ as a matrix circularly shifted by k columns. i.e.

$$I_{k,N} = \begin{bmatrix} 0 & \cdots & 0 & 1 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 1 \\ 1 & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix}.$$

Denote the following circulant matrices,

$$H = Cir[h(0), h(1)]_{N\times N} \text{ and } G = Cir[g(0), g(1)]_{N\times N}$$

Property 1. *Let a matrix A_0 denote an initial image. Its redundant representation using a separable Haar function (i.e., obtaining the following spectral decomposition Low-Low,Low-High, High-Low, High-High) can be written as*

$$A_1 = HA_0H'; \quad D_1 = HA_0G';$$

$$D_2 = GA_0H'; \quad D_3 = GA_0G',$$

where "$'$" denotes transposition. The reconstruction matrices can similarly be written as

$$R_1^h = h(0)I; \quad R_1^g = g(0)I;$$

$$R_2^h = h(1)I_{1,N}; \quad R_2^g = g(1)I_{1,N}$$

In light of the fact that a redundant representation is given or may be computed, the exact reconstruction methods have to be carefully rewritten. Towards that end we have the following:

Property 2. *Denoting the partial reconstruction matrices by*

$$\begin{aligned} RA_0^{ij} &= R_i^h A_1 R_j^{h\prime}; RD_1^{ij} = R_i^h D_1 R_j^{g\prime}; \\ RD_2^{ij} &= R_i^g D_2 R_j^{h\prime}; RD_3^{ij} = R_i^g D_3 R_j^{g\prime}, \quad i,j = 1,2. \end{aligned} \tag{5.6.1}$$

we may use any of the following four methods to exactly reconstruct the original image A_0*.,*

method1: $A_0 = RA_0^{11} + RD_1^{11} + RD_2^{11} + RD_3^{11}$

method2: $A_0 = RA_0^{21} + RD_1^{21} + RD_2^{21} + RD_3^{21}$

method3: $A_0 = RA_0^{12} + RD_1^{12} + RD_2^{12} + RD_3^{12}$

method4: $A_0 = RA_0^{22} + RD_1^{22} + RD_2^{22} + RD_3^{22}$.

5.7 Smoothing in the Frame Domain

The above decomposition and reconstruction procedures follow similar steps for other higher order wavelets such as Daubechies'. Accounting for the impulse response of corresponding filters leads to a a slight modification reflected in the matrices H and G which can be written as

$$\begin{aligned} H &= Cir[h(0), h(1), h(2), h(3)]_{N\times N}, \\ G &= I_{2,N} * Cir[g(-2), g(-1), g(0), g(1)]_{N\times N} \\ R_i^h &= I_{2,N} * Cir[h(i+1), 0, h(i-1)]_{N\times N}, i = 1,2 \\ R_i^g &= Cir[g(i-1), 0, g(i-3)]_{N\times N}, i = 1,2. \end{aligned}$$

Following the same strategy for Daubechies' wavelets as above, a reconstruction in a frame may be obtained, and any of the following representations may be used

$$\begin{aligned} A_0 &= R_1^h A_1 R_1^{h\prime} + R_1^h D_1 R_1^{g\prime} + R_1^g D_2 R_1^{h\prime} + R_1^g D_3 R_1^{g\prime}, \\ A_0 &= R_2^h A_1 R_1^{h\prime} + R_2^h D_1 R_1^{g\prime} + R_2^g D_2 R_1^{h\prime} + R_2^g D_3 R_1^{g\prime}, \\ A_0 &= R_1^h A_1 R_2^{h\prime} + R_1^h D_1 R_2^{g\prime} + R_1^g D_2 R_2^{h\prime} + R_1^g D_3 R_2^{g\prime}, \\ A_0 &= R_2^h A_1 R_2^{h\prime} + R_2^h D_1 R_2^{g\prime} + R_2^g D_2 R_2^{h\prime} + R_2^g D_3 R_2^{g\prime}. \end{aligned}$$

We next denote the detail matrix coefficients at the first level by $D_i, i = 1, 2, 3, 4$ and at the second level by $W_i^j, j = 1, 2, 3, 4$. Armed with methods 1-4 to reconstruct D_1, D_2, D_3, D_4, and using the knowledge that noise primarily dominates higher spectral bands, we proceed to effect the smoothing similar to that of a Haar frame-based linear diffusion (i.e., progressive elimination of detail of detail information from A_0) to result in the following recursion,

$$\begin{aligned} U_n &= U_{n-1} - \frac{1}{12}(R_1^h R_2^g W_1^3 R_2^{g'} R_1^{g'} \\ &+ R_1^g R_2^g W_2^3 R_2^{g'} R_1^{h'} + R_1^g R_2^g W_3^3 R_2^{g'} R_1^{g'}). \end{aligned} \tag{5.7.1}$$

Note that this recursion will also achieve a linear diffusion as stated in Proposition 1, albeit with modified intermediate characteristics. The complete smoothing witnessed with the linear Heat equation will still be the ultimate fate of the signal being filtered. A technique to slow down such an event is described next.

5.8 Nonlinear Reconstruction

Inspired by the algorithms of the first section such as that of Perona-Malik's or that proposed in [52] and to better address the preservation of features, such as texture which, however and as just shown, is eventually swept away by a linear diffusion. These features as noted above, are well captured by the correlation among the coefficients, which by using the insight of Section 3, help us proceed to construct a frame-based nonlinear reconstruction filter. The flexibility in properly selecting a wavelet function adapted to the texture of interest, together with the rationale of preserving large magnitude coefficients which best summarize the underlying information while reducing/eliminating the contribution of others as suggested by Eq. (5.7.1), lead us to propose a transformation of the individual coefficients as

$$D_i = D_i * \mathcal{N}(\{D_j\}). \tag{5.8.1}$$

The generally nonlinear functional may take a monotonic form similar to that proposed by P-M, where the decay rate is selected based on some prior knowledge we may have about the underlying image.

For illustrative purposes, we choose $\mathcal{N}(y) = e^{-\frac{y^2}{2K}}$, and hasten to point out that other

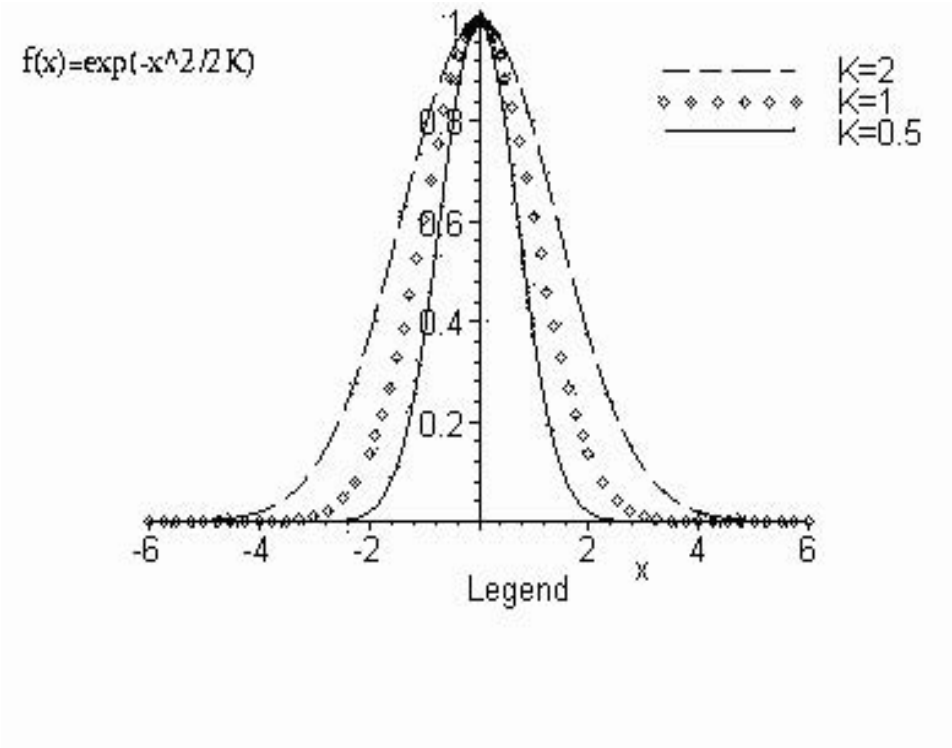

Figure 5.5: One possible nonlinear functional is an exponential weighting.

functionals adapted to other specific applications are currently under investigation. The set of coefficients which are subjected to the transformation are,

$$\begin{cases} D_1 = D_1 * exp(-D_1^2/2K); \\ D_2 = D_2 * exp(-D_2^2/2K); \\ D_3 = D_3 * exp(-D_3^2/2K), \end{cases}$$

and their insertion in the above recursive reconstruction yields a nonlinear filter.

5.9 Experimental Results

The absence in our illustrations of blocky artifacts or Gibbs phenomena so common with many multiscale techniques (wavelet thresholding) and also robust scale space techniques (e.g. [52]), not only demonstrates the effectiveness of the proposed approach, but also points to the importance of the synergy that may be gleaned from multiscale analysis and scale space methods. The performance of our proposed nonlinear filter, is readily assessed in the Lenna picture shown for three different denoising techniques, namely, our originally proposed technique[49], Perona-Malik's, and the newly proposed technique. The ability of the proposed technique to remove noise while preserving features like texture is readily apparent in Figures 5.6- 5.7 and the importance of such techniques in many applications needs no further elaboration.

Figure 5.6: A noisy Lenna image and filtered result with three algorithms.

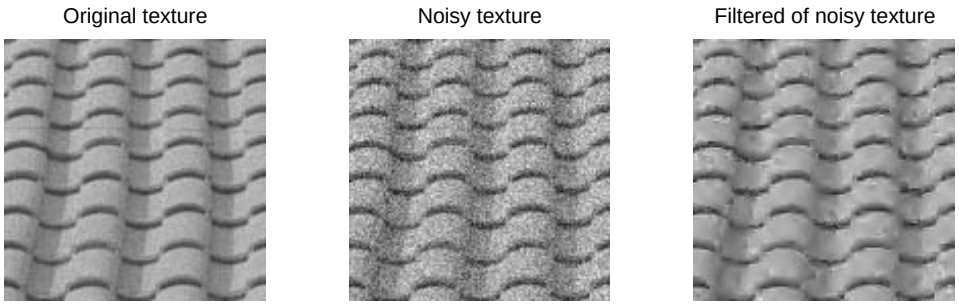

Figure 5.7: A texture image, noisy texture image and filtered result with Daubechies 4

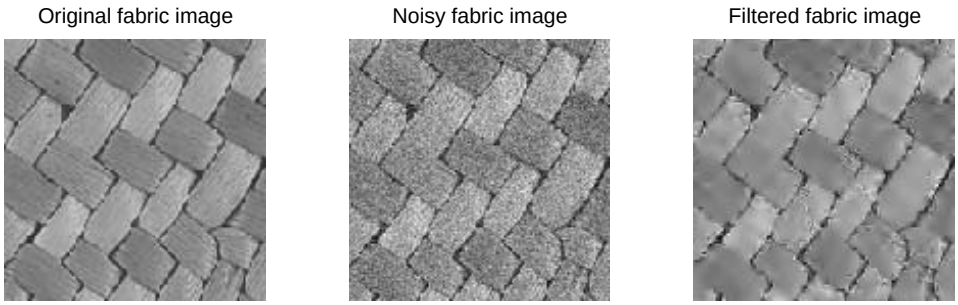

Figure 5.8: A texture image, noisy fabric image and filtered result with Daubechies 4

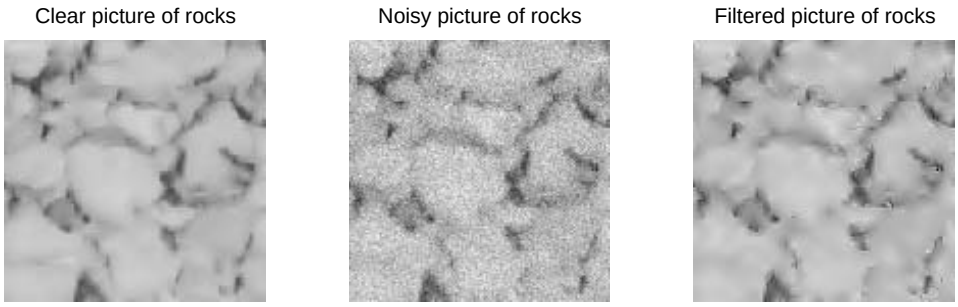

Figure 5.9: A texture with rocks image, its noisy image and filtered result with Daubechies 4

5.10 Appendix

Proof of Proposition 3.1: Having established the above two lemmas and specializing the results to a Haar function,

$$\phi(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{others,} \end{cases} \quad \psi(x) = \begin{cases} -1 & 0 < x < 0.5 \\ 1 & 0.5 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

for which the wavelet filter impulse response is $g(0) = -\frac{1}{\sqrt{2}}$, $g(1) = \frac{1}{\sqrt{2}}$, and which readily yields the following detail coefficient of $U(n, x)$

$$U^d(n, x) = -\frac{1}{\sqrt{2}}U(n, x) + \frac{1}{\sqrt{2}}U(n, x + 1).$$

Towards establishing the iterative implementation of the linear diffusion, we first invoke Eq.(5.4.3) of Lemma 1, to write

From Lemma 1 and we have

$$U(n, 2x) = h(0) \cdot U^a(n, 2x) + g(0) \cdot U^d(n, 2x). \tag{5.10.1}$$

To reconstruct $U^d(n, x)$ from its Haar wavelet decomposition coefficients $U^{ad}(n, x), U^{dd}(n, x)$, we use Eq.(5.4.4) to obtain

$$U^d(n, 2x) = h(1) \cdot U^{ad}(n, 2x - 1) + g(1) \cdot U^{dd}(n, 2x - 1). \tag{5.10.2}$$

Combining Eq.(5.10.1) and Eq.(5.10.2) results in the following

$$\begin{aligned} & U(n, 2x) \\ = \; & h(0) \cdot U^a(n, 2x) + g(0) \cdot (h(1) \cdot U^{ad}(n, 2x - 1) + g(1) \cdot U^{dd}(n, 2x - 1)) \\ = \; & h(0) \cdot U^a(n, 2x) + g(0) \cdot h(1) \cdot U^{ad}(n, 2x - 1) + g(0) \cdot g(1) \cdot U^{dd}(n, 2x - 1). \end{aligned} \tag{5.10.3}$$

We may similarly obtain

$$\begin{aligned} & U(n, 2x + 1) \\ = \; & h(0) \cdot U^a(n, 2x + 1) + g(0) \cdot h(1) \cdot U^{ad}(n, 2x) + g(0)g(1) \cdot U^{dd}(n, 2x), \end{aligned} \tag{5.10.4}$$

implying the following

$$\begin{aligned} & U(n, x) \\ = \; & h(0) \cdot U^a(n, x) + g(0) \cdot h(1) \cdot U^{ad}(n, x - 1) + g(0)g(1) \cdot U^{dd}(n, x - 1). \end{aligned} \tag{5.10.5}$$

Note that the second level detail component $g(0)g(1)U^{dd}(n, x-1)$ is equal to $-\frac{1}{2}U^{dd}(n, x-1)$, and is subtracted from the reconstruction of $U(n, x)$ to yield the updated $U(n+1, x)$. The following formula, equivalent to a Heat diffusion equation, is hence established,

$$\begin{aligned} U(n+1, x) &= U(n, x) - (-\frac{1}{2}U^{dd}(n, x-1)) \\ &= U(n, x) + \frac{1}{2}U^{dd}(n, x-1) \end{aligned} \tag{5.10.6}$$

■

Chapter 6

Independent Component Analysis

Independent component analysis(ICA), a data analysis concept that was first introduced by C.Jutten and J.Herault [47] and also known as blind source separation(BSS) in applications, has been of intense research interest in a number of application fields (for instance, speech recognition, remote sensing and biomedical imaging). Much has been accomplished [16, 14, 13, 42, 7, 76, 2] including the ICA demonstrated potential in signal/image enhancement. It may be viewed as a natural extension to standard principal component analysis(PCA), which, as is well known, is based on the correlation structure of observed data. In this chapter, we investigate novel and efficient ways of carrying out an ICA using novel information theoretic criteria.

6.1 Linear ICA models

Consider a set of observed signals from multiple sensors, each sensor receiving a different combination of the source signals, which, for simplicity, are assumed to be random variables, since if viewed as sample paths of a random process, more complex models are required. Hence, representing data as random vectors, as we elaborate, facilitates the use of statistical methods such as entropy, correlation and measurement of redundancy, and turn out to be a powerful model. Although a general data model, such as, $\boldsymbol{x} = f(\boldsymbol{s})$ is desirable, here $\boldsymbol{x} = (X_1, \cdots, X_m)$ are outputs, $\boldsymbol{s} = (S_1, \cdots, S_n)$ is a source random vector and f is a

transform function, most applications assume a linear transform in the form of $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$. To proceed with describing a linear ICA model, we will assume that

- The source signals are independent random variables
- The distributions of the source signals are unknown

The task in using of independent component analysis(ICA) or Blind source separation(BSS) is to recover independent source signals from mixed observations under these assumptions.

The linear BBS or ICA model assumes the existence of n independent source signals $\boldsymbol{s}(k) = (S_1(k),\ S_2(k),\ \cdots,\ S_n(k))$, and $\boldsymbol{x}(k) = (X_1(k),\ X_2(k),\ \cdots,\ X_m(k)), k = 1, \cdots, K$ are observed signal samples from m sensors, which are linear mixtures of source signals in the presence of additive noise $\boldsymbol{n}(k) = (N_1(k),\ N_2(k),\ \cdots,\ N_m(k))$, and more simply,

$$\boldsymbol{x}(k) = \boldsymbol{A}\boldsymbol{s}(k) + \boldsymbol{n}(k)$$

where $\boldsymbol{A}$ is an unknown full column rank $m \times n$ matrix that accounts for the linear mixtures of the signals, with K observed discrete samples. In this model, $\boldsymbol{s}(k)$ and $\boldsymbol{x}(k)$ could be complex signals and $m \geq n$ is usually imposed (some special $m < n$ cases have been studied in [17, 59, 10]), however, we only consider the simplest model, namely, real-valued signals, $m = n$, and noise free observations, i.e.,

$$\boldsymbol{x}(k) = \boldsymbol{A}\boldsymbol{s}(k),$$

since this model captures the essence of the ICA or BSS problem, where noise is usually considered as a nuance parameter and is largely neglected in the literature. An example of mixed signals obtained from independent sources subjected to rotation is displayed in Fig. (6.1).

To recover signals from mixed data is to estimate a full rank matrix $\boldsymbol{W}$ such that the estimated signal components of $\boldsymbol{y}(k) = (Y_1(k),\ \cdots,\ Y_n(k))$ are as independent as possible and the outcome sources space (may be permuted and scaled) are close to source signals, namely

$$\boldsymbol{y}(k) = \boldsymbol{W}\boldsymbol{x}(k) = \boldsymbol{W}\boldsymbol{A}\boldsymbol{s}(k) = \boldsymbol{B}\boldsymbol{s}(k).$$

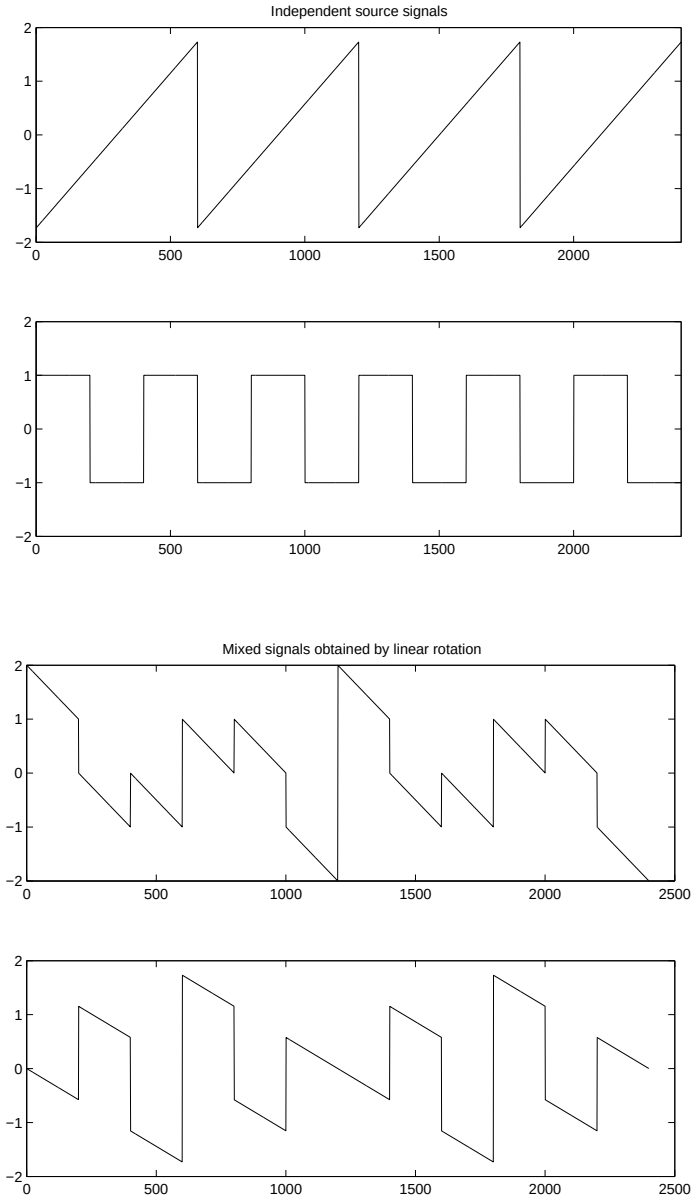

Figure 6.1: Independent source signals and mixed signals obtained by rotation.

We assume that at most one of the source signals $S_i(k)$ is allowed to have a Gaussian distribution. This follows from the fact that it is impossible to separate several Gaussian sources from each other [16].

6.2 Existing ICA Algorithms

ICA algorithms are closely related to principal component analysis [46] , factor analysis [36], and projection pursuit algorithms [30, 41]. In many ICA algorithms, it is required that mixture data be normalized by its variance and be pre-whitened, which means that we can always assume a unity variance for each component (we also assume 0 mean-valued data, as in practice, we can always subtract the mean value from the original data) and that there exists a whitening transform matrix $\boldsymbol{V}$ such that the whitened data $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{X}$ has a correlation matrix $\boldsymbol{R} = E(\boldsymbol{U}\boldsymbol{U}^\tau) = \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix.

Further, by assuming that at most one of the components is Gaussian distributed, and a full rank matrix A, which together with the fact that independence is not affected by different ordering and scaling of the components, a unique solution to ICA is assured in the sense that permutation and scaling are allowed for each component.

Upon observing the mixed signals $\boldsymbol{x} = (X_1, \cdots, X_n)$, many algorithms seeking to estimate the matrix $\boldsymbol{W}$ have been proposed on the basis of information theoretic as well as neural network-based criteria. These algorithms consist of optimizing proposed objective functions known as contrast functions, or cost functions [16]. Valid contrast functions must be designed in such a way that the source separation is achieved when they reach their optimal (minimum or maximum) values. Common contrast functions are based on measures such as, entropy, high-order cumulants, divergence among the joint probability density functions of observed data for independent models.

The first neural source separation algorithm was presented in [47], and has used the principle of cancelling non-linear cross-correlations of the form $E\{g_1(Y_i)g_2(Y_j)\}$ (where $g_1,\ g_2$ are some suitably chosen odd non-linear functions, and $Y_i,\ Y_j$ are estimations of source independent signals) to achieve independent components, which in turn implies that $E\{g_1(Y_i)g_2(Y_j)\} = E\{g_1(Y_i)\}E\{g_2(Y_j)\}$ when $Y_i,\ Y_j$ are independent. Further, a nonlinear objective function

based on a generalization of PCA is proposed by introducing a nonlinear function $g(x)$ to seek out principal components by (see [44, 70]),

$$\boldsymbol{w}_1 = \arg \max_{||\boldsymbol{w}||=1} E\{(g(\boldsymbol{w}^\tau \boldsymbol{x}))^2\}. \tag{6.2.1}$$

Inspired by neural networks, a contrast function was derived in [7] on the basis of maximizing the Shannon entropy of non-linear outputs from a neural network. Specifically, one is to maximize the following formula

$$L = H(g_1(\boldsymbol{w}_1^\tau \boldsymbol{x}),\ \cdots, g_m(\boldsymbol{w}_m^\tau \boldsymbol{x})), \tag{6.2.2}$$

where $\boldsymbol{x}$ is the input to a neural network whose outputs are of the form $g_i(\boldsymbol{w}_i^\tau \boldsymbol{x})$ and $\boldsymbol{w}_i$ are the weight vectors of the neurons. Here Shannon differential entropy of a continuous random variable (or a random vector) X with a probability density function (pdf) $f(x)$ is defined as

$$H(X) = -\int_\Omega f(x) \log f(x) dx. \tag{6.2.3}$$

In [16], it was proposed to use mutual information as a contrast function to more fully describe the statistical independence property. In a probabilistic/information theoretic setting, researchers have held the mutual information measure as one of choice in identifying a proper representation basis of random samples of signals/sources, and for which the resulting coefficients are independent. This is an interesting property of source independence as it does not include any implicit or explicit assumption about the distributions of the sources. Mutual information for a set of random variables $(Y_1,\ \cdots,\ Y_n)$ may be written as

$$I(Y_1,\ \cdots,\ Y_n) = \sum_{i=1}^{n} H(Y_i) - H(Y_1,\ \cdots,\ Y_n) \tag{6.2.4}$$

Using the maximum likelihood principle and, if we assume a source vector $\boldsymbol{s}$ is distributed according to a pdf $q(s)$, a contrast function is formulated as the log-likelihood function of a pdf of the estimated output $\boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{x}$, for T samples $\boldsymbol{x}(1), \cdots, \boldsymbol{x}(T)$, and output $\boldsymbol{y}(1), \cdots, \boldsymbol{y}(T)$, to take the form [13, 18]:

$$\phi_{ML}^T(\boldsymbol{y}) = \log p(\boldsymbol{y}) = \frac{1}{T}\sum_{k=1}^{T} \log q(\boldsymbol{A}^{-1}\boldsymbol{x}(k)) - \log(det(\boldsymbol{A})). \tag{6.2.5}$$

By the law of large numbers, when $T \to \infty$, the log-likelihood function converges to the expectation of $\log q(\boldsymbol{A}^{-1}\boldsymbol{x}) + constant$, denoted as $\phi_{ML}(\boldsymbol{y}) = -E\{\log q(\boldsymbol{A}^{-1}\boldsymbol{x})\}$. This may also be motivated by the Kullback-Leibler divergence $K(f|g)$, which can be viewed as a distance between two probability density functions f and g (though it is not a real distance measure because it is not symmetric) and is also written as $K(X|Y)$ where f and g are pdfs of two random variables X and Y. The definition of Kullback-Leibler divergence [53] is the following

$$K(f|g) \triangleq \int_S f(s) \log\left(\frac{f(s)}{g(s)}\right) ds. \tag{6.2.6}$$

This in turn may be used to establish that the log-likelihood function $\phi_{ML}(\boldsymbol{y})$ may be written as

$$\phi_{ML}(\boldsymbol{y}) = K(\boldsymbol{y}|\boldsymbol{s}). \tag{6.2.7}$$

The Maximum Likelihood principle thus attempts to find a matrix A such that the distribution of $\boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{x}$ is as close as possible to the hypothesized joint distribution of the sources. Denote $\tilde{\boldsymbol{y}}$ as a random vector with independent entries, and each entry distributed as the corresponding entry of $\boldsymbol{y}$ ($\tilde{\boldsymbol{y}}$ is thus called the factorial distributed random variable of Y), we see that

$$K(\boldsymbol{y}|\boldsymbol{s}) = K(\boldsymbol{y}|\tilde{\boldsymbol{y}}) + K(\tilde{\boldsymbol{y}}|\boldsymbol{s}), \tag{6.2.8}$$

for any vector with independent entries. Eq. (6.2.8) shows that the minimum value of $K(\boldsymbol{y}|\boldsymbol{s})$ is reached by minimizing both terms,

1. the first right term of Eq. (6.2.8), which is, in fact, mutual information, therefore required independence among each entry of $\boldsymbol{y}$

2. the second right term of Eq. (6.2.8), which requires that the individual entries of $\boldsymbol{y}$ have the same distribution as those of $\boldsymbol{s}$.

we can thus see that, if the source distributions are known, ϕ_{ML} is more accurate because it expresses directly the fitness between data and model. The initial distributions of the source signals are however, usually unknown, which reduces the maximum likelihood algorithm to that of maximizing mutual information by only considering an independence property, which of course remains an important feature of the sources.

It is also proved in [13] that the principle of neural network entropy maximization, namely, the infomax principle, is equivalent to maximum likelihood estimation by properly choosing a function $g = (g_1, \cdots, g_m)$ in Eq. (6.2.2) as a cumulative distribution function (cdf) of the source variables, and if known.

While the above measures are sound and theoretically appealing, their big drawback is in having to estimate the pdf's or the entropy which is not always trivial. Several approximations to MI based on polynomial Taylor density expansions have been proposed and yielding contrast functions based on higher order cumulants [16, 12]. Cumulants of order 2 and 4 have been predominantly used and yielding for example the following approximations of $\boldsymbol{y} = (Y_1, \cdots, Y_n)$ of the form [16],

$$I(\boldsymbol{y}) \approx C + \frac{1}{48}\sum_{i=1}^{n}[4k_3(Y_i)^2 + k_4(Y_i)^2 + 7k_4(Y_i)^4 - 6k_3(Y_i)^2k_4(Y_i)] \tag{6.2.9}$$

where $k_i(X) = E(X^i),\ i = 1, 2, \cdots$, are $i-th$ order cumulants of a random variable X and C is a constant [16]. The approximation, however, is valid only when the density function of Y is close to the Gaussian probability density function, otherwise, poor estimate may result.

Higher-order cumulant tensors have also been directly used as criteria by taking advantage of the prevailing algebraic structure [9, 10, 11], a good review may be found in [12]. This method consists of looking for the eigenvectors of a higher order cumulant tensor. The fourth-order cumulant tensor can be defined as the following linear operator $\boldsymbol{T}$ from the space of $m \times m$ matrices to the space of $m^2 \times m^2$ matrices with the i, j element of $\boldsymbol{T}$ as:

$$\boldsymbol{T}(\boldsymbol{K})_{ij} = \sum_{k,l} cum(X_i,\ X_j,\ X_k,\ X_l)\boldsymbol{K}_{kl} \tag{6.2.10}$$

where $cum(X_i,\ X_j,\ X_k,\ X_l)$ denotes the fourth-order cumulant and the subscript ij means the $(i,\ j)-th$ element of a matrix, and $\boldsymbol{K}_{kl}$ is a $m \times m$ matrix. This linear operator has m^2 eigenvalues. Solving for the eigenvectors of this eigenmatrix would lead to an estimation of the ICA model. The advantage of this method is that it requires no knowledge of the distribution of the independent source components, with the understanding that the efficiency issue constraints it to small dimensions.

With emergence of approximation techniques of differential entropy, more sophisticated

approximations of mutual information may be constructed and applied to ICA [43, 22]. Non-parametric estimation of mutual information [19] based on dependent data also provides a useful technique to directly implement ICA algorithms and further motivating the investigation of alternative measurements, such as Jensen-Rényi divergence [37, 1] and Rényi mutual divergence as criteria to ICA [5]. The technique developed to approximate $\alpha-$ Rényi entropy and Rényi divergence are also described in [5, 1, 38, 39].

Practical consideration, such as an unreliable faulty source, or that the underlying sources may in fact be Gaussian, may lead one to favor a technique which would avoid over-assumptions about the prevailing statistics over another with selectable data ranges (e.g. ignore large outliers) and with varying degrees of weighted contributions instead of the commonly used uniform equal weighting as is typical of many existing measures, such as weighted covariance matrix.

Miscellaneous alter existing approaches are deferred to the literature [76], [2] and references therein as we have instead focused on techniques relevant to our later discussing.

6.3 Applications of ICA

ICA, or BBS techniques have been applied in any fields where an array of m receivers collect data of linear mixtures of n source signals. Examples include speech separation (known as 'cocktail party problem') as several microphones are placed in different points while there are several speakers. It may be also applied in processing arrays of radar or sonar signals and processing of multi-sensor biomedical recording signals, such as EEG, MEG signals used to record brain activities. Medical imaging such as fMRI, processing of geophysical data and restoration of image features are also other typical applications.

Chapter 7

New measure criteria for ICA

As we mentioned earlier, mutual information(MI), as a measure between two probability densities has been intensively used by many authors as a contrast function for ICA. Its estimation is complicated by having to use empirical density functions, which, results in a weakness to estimate entropies. Research has mainly focuses on finding higher-order approximations of mutual information or different techniques. This criterion, namely, mutual information, however, has recently been fallen out of favor to other information measures, such as $\alpha-$Jensen-Rényi ($\alpha-$JR) divergence [37]. This divergence measure provides a distance among a group of probability densities, and thus can serve as an alternative and improved criterion to ICA. Also with the increasing interest in simplified and robust estimations of mutual information, some recent results [1, 19] are significant enough to allow us to reasonably consider applying them in ICA investigations. In this chapter, we propose Rényi mutual divergence as a new criterion on account of its additional features over mutual information. We also propose a technique to approximate Rényi mutual divergence by analyzing dependent data, and discuss in detail its properties later in this chapter.

7.1 Definition of Rényi Entropy

To introduce two divergence measures based on the Rényi entropy, an information measure that was first introduced by Rényi in [81, 82], which has been shown to be theoretically as

well as practically useful, we provide a brief overview on $\alpha-$Rényi entropy.

For a discrete random variable X with sample space Ω, whose corresponding probability distribution $P(x_i) = P(X = x_i),\ x_i \in \Omega,\ i \in I$, where I is an index set, and for $\alpha \in (0, 2],\ \alpha \neq 1$, the $\alpha-$Rényi Entropy is defined as

$$H^{\alpha}(P(x)) = \frac{1}{1-\alpha} \ln \left(\sum_{i \in I} P^{\alpha}(x_i) \right), \tag{7.1.1}$$

with "ln" denoting the natural logarithm.

For a continuous random variable X with a probability density function (pdf) $p(x),\ x \in \mathcal{R}^d$, $\alpha-$Rényi entropy is defined as

$$H^{\alpha}(p(x)) = \frac{1}{1-\alpha} \ln \left(\int p^{\alpha}(x) dx \right), \tag{7.1.2}$$

where $\alpha \in (0, 2],\ \alpha \neq 1$.

We can see that, when $\alpha \to 1$, $\alpha-$Rényi entropy degenerates to the Shannon entropy [85], which is defined as

$$H(p(x)) = -\int p(x) \ln p(x) dx, \tag{7.1.3}$$

for a continuous random variable and

$$H(P(x)) = -\sum_{i \in I} P(x_i) \ln P(x_i), \tag{7.1.4}$$

for a discrete random variable.

The advantage of Rényi entropy is that the probability (or the probability density) is modulated by a factor α, which along with its simpler form make Rényi entropy easier to implement, and with better adapted statistical properties to a random variable. Rényi entropy is concave for all $\alpha \in (0, 2]$ and Fig. (7.1) demonstrates this property for a specific example with Bernoulli distributions for different $\alpha \in (0, 2]$.

7.2 Jensen-Rényi divergence as a new criterion for ICA

7.2.1 Introduction to $\alpha-$Jensen-Rényi divergence

$\alpha-$Jensen-Rényi ($\alpha-$JR) divergence is a new concept recently proposed in [37], and may be viewed as a generalization of Rényi entropy [81, 82] and of Jensen-information [61], see

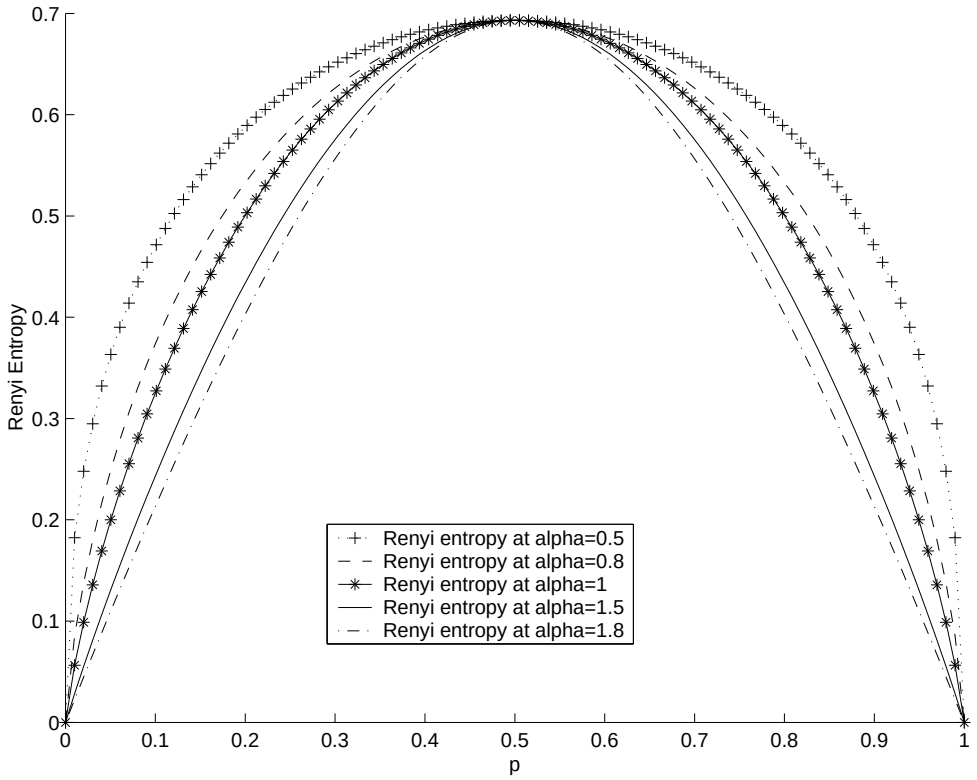

Figure 7.1: Renyi entropy of Bernoulli distributions at several α compared to Shannon entropy

[1] for a thorough investigation of Rényi entropy within the context of Minimum Spanning Trees(MST). The $\alpha-$JR divergence as an information measure, invokes a weighted combination of several distributions whose contribution is modulated by way of an α-exponentiation parameter (Viz. Eq. 7.2.1).

The generalized nature of this measure has shown very promising results in applications [37] and exhibits a wide scope of applicability tied to one's ability to reinterpret the population densities and to hence reexpress the measure itself. Specifically, this measure may simply be interpreted as one of *independence* between two or more probability density functions (data population) and is therefore naturally applicable to the well known problem of independent component analysis (ICA) (or also known as the *Blind Source Separation* problem).

In light of this practical interest and of the intrinsic properties of the $\alpha-$JR divergence measure, we propose it be the basis for a new criterion as further elaborated on in the next

subsection. In spite of its numerous advantages, our proposed technique nevertheless unveils a limitation in its computational implementation, where a bottleneck emerges in the course of estimating probability densities.

7.2.2 $\alpha-$Jensen-Rényi divergence as an Independence Measure

For a set of probability distributions $\{P_i(x)\}_{i=1\cdots n}$, $\alpha-$JR divergence is defined as

$$\begin{aligned} & JR_\alpha^{\{\omega_i\}_{i=1\cdots n}}(\{P_i(x)\}_{i=1\cdots,n}) \\ = \; & H^\alpha(\sum_{i=1}^n \omega_i P_i(x)) - \sum_{i=1}^n H^\alpha(P_i(x))\omega_i, \end{aligned} \tag{7.2.1}$$

with $\sum_i \omega_i = 1$, $0 \leq \omega_i \in \mathcal{R}$ and $0 < \alpha < 2$. The $\alpha-$JR divergence measures the distance between two or more distributions by adjusting weights on different distributions.

By Jensen's inequality, one can show that the above expression is minimized and achieves a 0-value if the distributions are identical, and is maximized when they are all different (i.e. each distribution function is a Dirac function positioned in different locations). If applied to a set of conditional distributions of X, Y, for example, a minimization of the $\alpha-$JR divergence would be tantamount to establishing that the distributions of X conditioned on Y, for all Y, are equal, hence implying the independence of X and Y. It is worth noting that this measure provides an additional flexibility of choosing α and ω_i, hence affording the selectivity among the data as mentioned earlier.

Given the essence of an ICA problem, the fitting measure to adopt is that of *independence*, which raises a natural question of how to reinterpret the $\alpha-$JR divergence to elicit such information contained in observed random variables from different populations. By defining in Eq. (7.2.1) $p_i(x) = P(X = x|Y = y_i)$ (i.e., as a conditional probability), we can easily conclude that, $JR_\alpha^{\{\omega_i\}_{i=1\cdots n}}(\{p_i(x)\}_{i=1\cdots n})$, denoted by $JR_\alpha(X, Y)$, yields a measure of independence between X and Y when it is minimized.

7.2.3 Application to ICA

To proceed with the description of the source separation problem, we denote the observations $X_i,\ i = 1\cdots n$, $\boldsymbol{x} = (X_1, X_2, \cdots, X_n)$ as a result of a mixing action of an unknown matrix

$\boldsymbol{A}$ on source data $\boldsymbol{s} = (S_1, S_2, \cdots, S_n)$, expressed as a sequence of independent random variables $S_i,\ i = 1 \cdots n$ of 0-mean and unit-variance, and more explicitly as a linear model

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}.$$

Our goal is to then recover $\boldsymbol{s}$ from merely observing $\boldsymbol{x}$. The solution to this problem requires one to typically first proceed to whiten the data, i.e., diagonalize the data covariance matrix so that $\boldsymbol{I}_n = \boldsymbol{W}\boldsymbol{A}\boldsymbol{A}^{\tau}\boldsymbol{W}^{\tau}$, where "$\tau$" denotes transposition and $\boldsymbol{I}_n$ is an identity matrix. The subsequent step is to search for an adapted pairwise rotation of axes to yield independent data along these directions, and effected by

$$\theta\theta_{ij}^{\{\omega_k\}_{k=1\cdots n}} = \arg\min_{\theta}(JR_{\alpha}^{\{\omega_i\}_{i=1\cdots n}}(X_i^{\theta}, X_j^{\theta}))$$

where $X_i^{\theta}, X_j^{\theta}$ are the corresponding random variables to X_i, X_j rotated by an angle θ. By iterating this processing to other pairs, all of the independent components are gleaned.

To illustrate the proposed technique, we provide two mixture cases: the first shown in Fig. (7.2), consists of two acoustic speech signals, and the second shown in Fig. (7.4) includes signals with heavy tail distributions. The recovered signals in both cases are shown in the corresponding figures, and demonstrate the effectiveness of the proposed technique. In Figs. (7.3) and (7.5) we compare and display the potential gains of using JR divergence over mutual information. The sharper and more significant nulls of the JR measure suggest a resilience and additional robustness in the presence of perturbations such as estimation errors.

It is important to note that the choice of parameters "ω_i" and "α" in this example have not been necessarily optimized (uniform prior chosen somewhat arbitrarily and "α" selected in light of the underlying signals (i.e. $1 < \alpha < 2$ for super-gaussian processes and $0 < \alpha < 1$ for sub-gaussian processes). This in fact, and to the best of our knowledge, remains an *open problem*.

7.3 $\alpha-$Rényi mutual divergence as a New Criterion for ICA

7.3.1 Introduction to $\alpha-$Rényi divergence

$\alpha-$Rényi divergence, an important information divergence introduced by Rényi [81, 82], can also be broadly viewed as a distance measure between two probability density functions in spite of its non-symmetric structure (it may be made symmetric at a cost of a more complex form), it is defined as

$$RD_\alpha(f,g) = \frac{1}{\alpha-1}\ln\left(\int f^\alpha(x)g^{1-\alpha}(x)dx\right), \tag{7.3.1}$$

where $0<\alpha<2, \alpha\neq 1$ and $f(x),\ g(x),\ x\in\mathcal{R}^d$ are two probability density functions of two random variables X and Y. This divergence shares the same maximization and minimization points as $\alpha-$JR divergence and KL-divergence, namely, the minimization point is reached when $f(x)=g(x)$, the maximization point is reached when $f(x), g(x)$ are totally different (or, one of the densities $f(x), g(x)$ should be a Dirac function in $\mathcal{R}^d$).

The definition of $\alpha-$Rényi divergence is fairly general in that some other divergences are its special case for a specific value of α. We can, for instance, see that KL-divergence is obtained when $\alpha\to 1$,

$$\lim_{\alpha\to 1} RD_\alpha(f,g) = \int f(x)\ln\frac{f(x)}{g(x)}dx, \tag{7.3.2}$$

and when $\alpha=1/2$, the so-called Battacharya distance is obtained

$$RD_{1/2}(f,g) = -2\ln\left(\int\sqrt{f(x)g(x)}dx\right). \tag{7.3.3}$$

As is well known, KL-divergence becomes the mutual information if we take $f(x)$ as a joint probability density of a random vector $\boldsymbol{x}=(X_1,\cdots,X_n)$, where $X_i,\ i=1,\cdots,n.$ is a random variable with $\mathcal{R}^d$ as a sample space, and $g(x)$ as the factorial probability density in the form of $g(x)=f_1(x)\cdots f_n(x)$, where $f_i(x),\ i=1,\cdots,n,\ x\in\mathcal{R}^d$ is the pdf of each individual random variable $X_i,\ i=1\cdots n$. When the two probability densities are used to write a $\alpha-$ Rényi divergence, it is called $\alpha-$Rényi mutual divergence, and is denoted as $MD_\alpha(\boldsymbol{x})$ or $MD_\alpha(X_1,\cdots,X_n)$. $\alpha-$Rényi mutual divergence measures the dependency among a group of random variables $X_1,\cdots,X_n$ by the distance between its joint pdf and its factorial pdf.

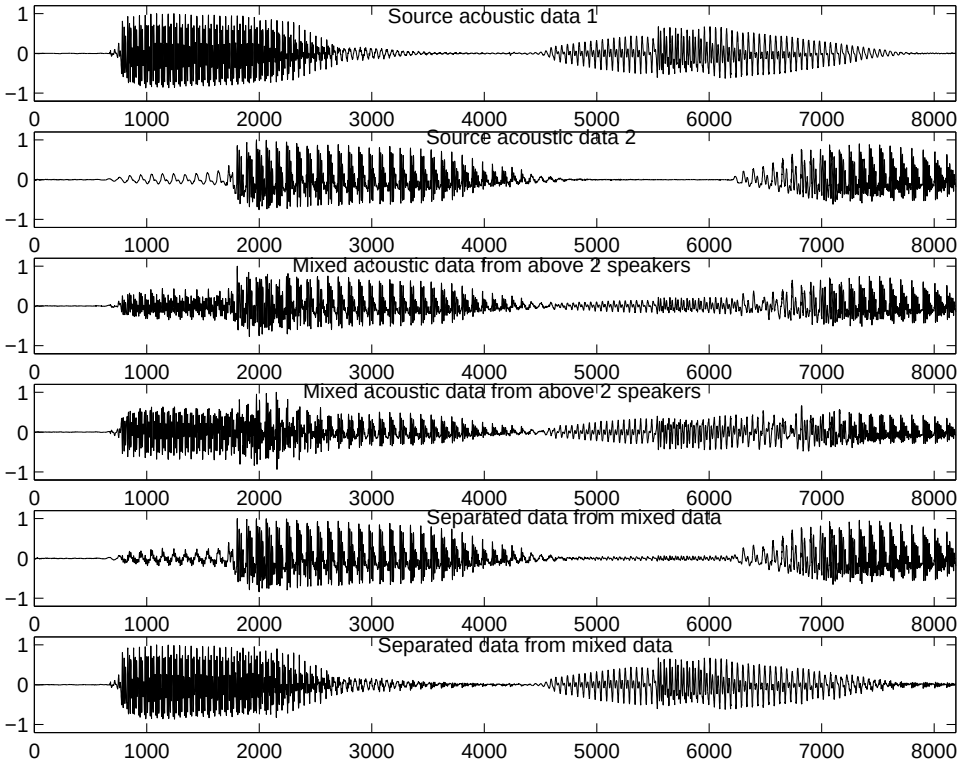

Figure 7.2: An application using α-JR Divergence

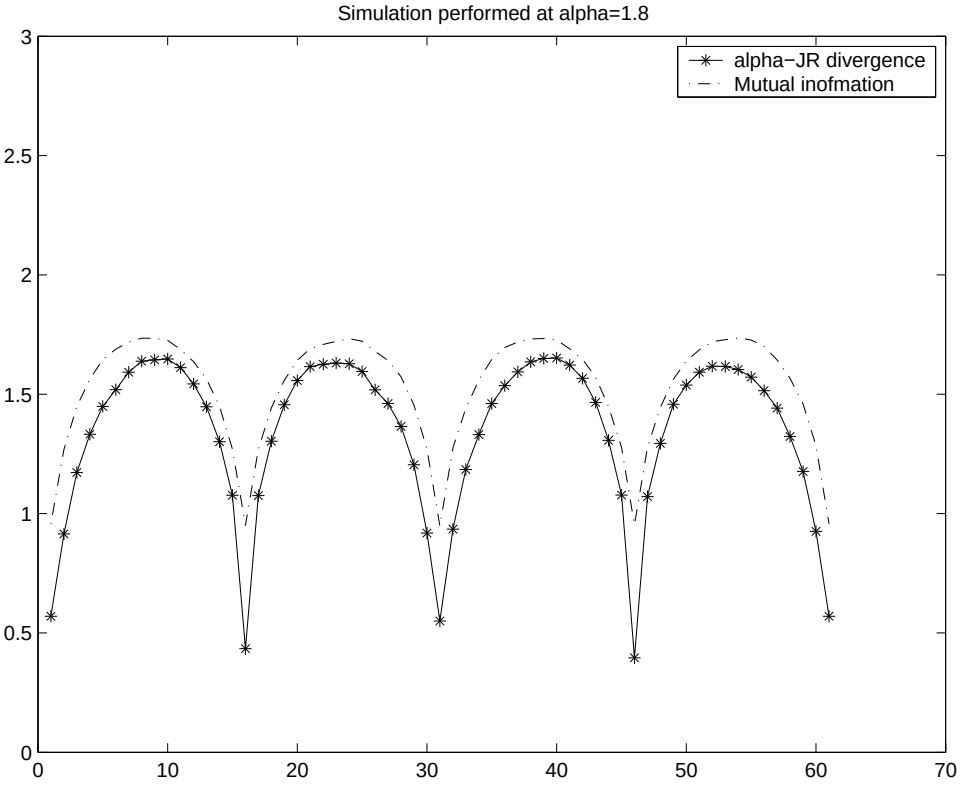

Figure 7.3: ICA criterion using mutual information and α-JR divergence.

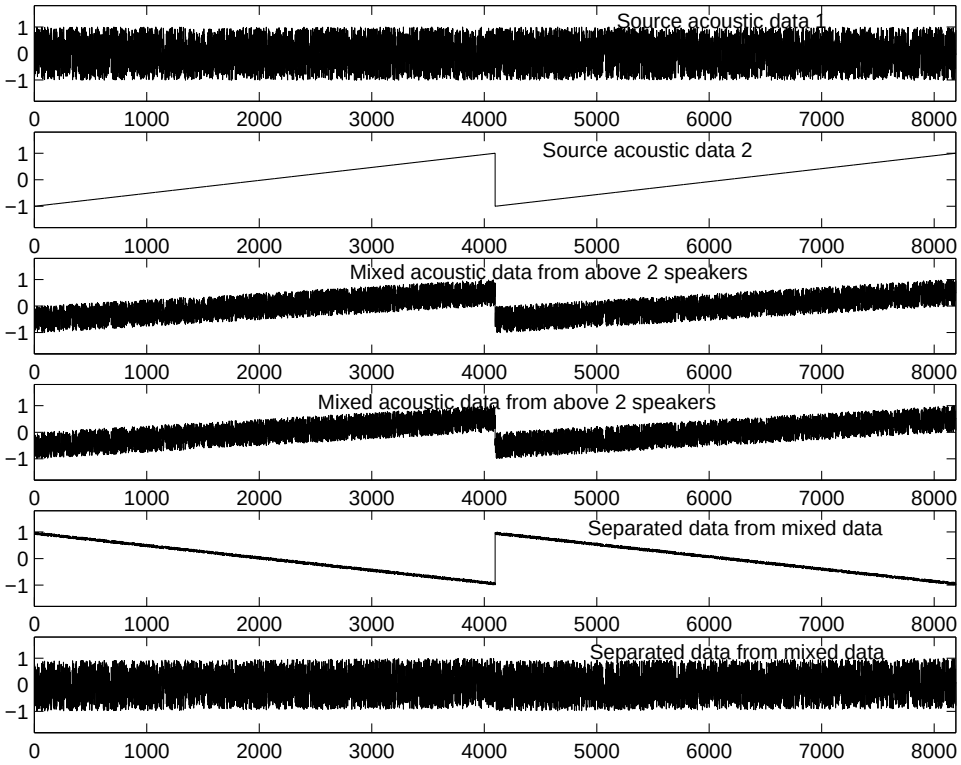

Figure 7.4: An application using α-JR Divergence

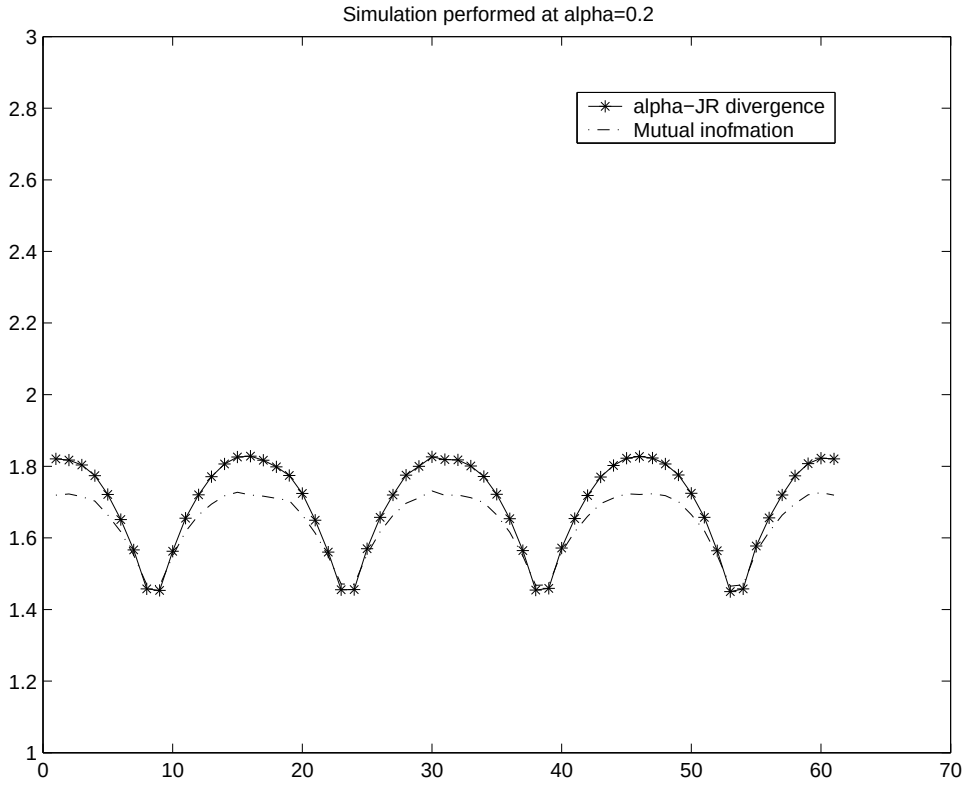

Figure 7.5: ICA criterion using mutual information and α-JR divergence.

Thus, $\alpha-$Rényi mutual divergence, as a generalization of mutual information, can serve as an independence information measure with its value depending on the index α applied to each distribution. In the following subsection, we restrict $\alpha-$Rényi mutual divergence to two random variables (X, Y) and prove a theorem that compares the degree of independence measurement of mutual information and Rényi mutual divergence between the two random variables. This therefore provides a theoretical argument for one to apply Rényi mutual divergence to ICA.

7.3.2 Comparison between $\alpha-$Rényi mutual divergence and Mutual Information

In order to illustrate the advantage of using $\alpha-$Rényi mutual divergence (with $1 < \alpha < 2$) as a criterion for ICA, we prove the following theorem,

Theorem 9. *For two continuous random variables X, Y with joint pdf $h(x, y)$ and marginal distributions $f(x)$, $g(y)$, we have the following inequality between $\alpha-$Rényi mutual divergence $MD_\alpha(X, Y)$ and mutual information $I(X, Y)$,*

case 1: Given $0 < \alpha < 1$,

$$MD_\alpha(X, Y) \leq I(X, Y)$$

case 2: Given $1 < \alpha \leq 2$,

$$MD_\alpha(X, Y) \geq I(X, Y)$$

Proof: For the joint probability density $h(x, y)$ of random variables X, Y and the corresponding marginal pdfs $f(x), g(y)$, the $\alpha-$Rényi mutual divergence may be written as

$$\begin{aligned} MD_\alpha(X, Y) &= \frac{1}{\alpha - 1} \ln \int h^\alpha(x, y) \left(f(x)g(y)\right)^{1-\alpha} dxdy \\ &= \frac{1}{\alpha - 1} \ln \int h(x, y) \left(\frac{h(x, y)}{f(x)g(y)}\right)^{\alpha-1} dxdy \\ &= \frac{1}{\alpha - 1} \ln E\left\{\left(\frac{h(X, Y)}{f(X)g(Y)}\right)^{\alpha-1}\right\} \end{aligned} \tag{7.3.4}$$

Since the function $p(x) = \ln(x)$ is strictly concave, according to Jensen-Inequality, we have

$$E\{p(X)\} \leq p(E\{X\})$$

from which, we have that, for two random variables X, Y,

$$E\left\{\ln\left(\frac{h(X,Y)}{f(X)g(Y)}\right)^{\alpha-1}\right\} \leq \ln E\left\{\left(\frac{h(X,Y)}{f(X)g(Y)}\right)^{\alpha-1}\right\} \tag{7.3.5}$$

when $0 < \alpha < 1$, we have $\alpha - 1 < 0$, thus

$$\begin{aligned} MD_\alpha(X,Y) &= \frac{1}{\alpha-1}\ln E\left\{\left(\frac{h(X,Y)}{f(X)g(Y)}\right)^{\alpha-1}\right\} \\ &\leq \frac{1}{\alpha-1}E\left\{\ln\left(\frac{h(X,Y)}{f(X)g(Y)}\right)^{\alpha-1}\right\} \\ &= E\left\{\ln\left(\frac{h(X,Y)}{f(X)g(Y)}\right)\right\} \\ &= I(X,Y) \end{aligned} \tag{7.3.6}$$

when $1 < \alpha \leq 2$, we have $\alpha - 1 > 0$, thus

$$\begin{aligned} MD_\alpha(X,Y) &= \frac{1}{\alpha-1}\ln E\left\{\left(\frac{h(X,Y)}{f(X)g(Y)}\right)^{\alpha-1}\right\} \\ &\geq \frac{1}{\alpha-1}E\left\{\ln\left(\frac{h(X,Y)}{f(X)g(Y)}\right)^{\alpha-1}\right\} \\ &= E\left\{\ln\left(\frac{h(X,Y)}{f(X)g(Y)}\right)\right\} \\ &= I(X,Y) \end{aligned} \tag{7.3.7}$$

Noted that $' ='$ is established only when $h(x,y) = f(x)g(y)$, namely X, Y are independent, which concludes the proof. ■

We here give two examples of this theorem with value $\alpha = 0.2$ and $\alpha = 1.8$, Fig.(7.6) and Fig(7.7). These two pictures compare the theoretical values of $\alpha-$Rényi mutual divergence of a sequence of Gaussian sources with that of mutual information, the formula of the $\alpha-$Rényi mutual divergence of a sequence of Gaussian sources is calculated in Section 7.5, we also draw the approximations of $\alpha-$Rényi mutual divergence and mutual information in

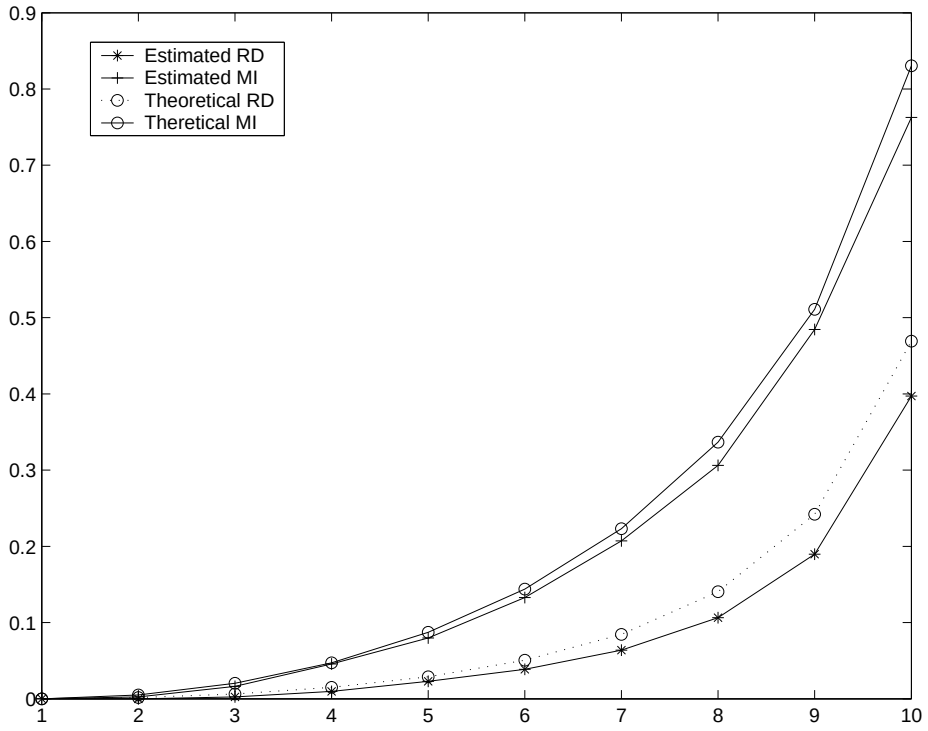

Figure 7.6: Approximated 0.2−Rényi mutual divergence and its exact theoretical value compared to mutual information

these two pictures, again, detail knowledge about the approximation is deferred to Section 7.5.

7.4 Application to ICA

As a result of the above theorem, we can see that when $1 < \alpha \leq 2$, $\alpha-$Rényi mutual divergence is a better adapted measure. This is made more compelling when considering the

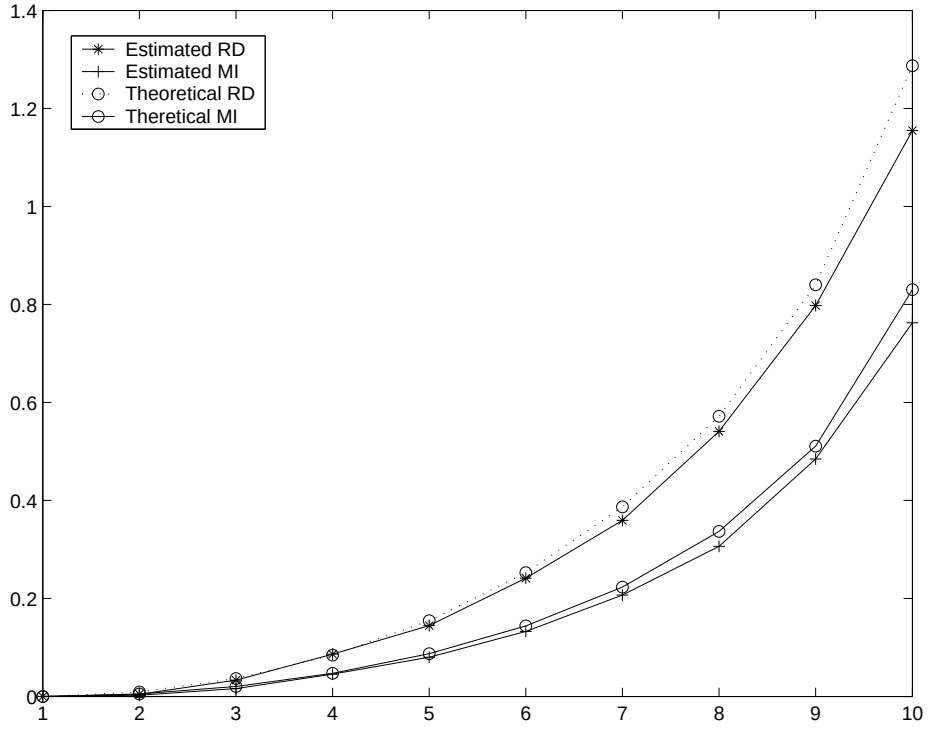

Figure 7.7: Approximated 1.8−Rényi mutual divergence and its exact theoretical value compared to mutual information

fact that $\alpha-$Rényi divergence is a concave function in each distribution, and that $\alpha-$Rényi mutual divergence attains a 0 minimum when two random variables are independent, and a maximal $H^{2-\alpha}(X)$ when the two random variables are fully dependent. In our application of such a measure, we propose a non-parametric estimation of $\alpha-$Rényi mutual divergence based on dependent data as shown in section 7.5.

The target function we use for ICA is the following

$$\theta = \arg\min_{\theta}(MD_{\alpha}(X_i^{\theta}, X_j^{\theta}))$$

The technique we use here is similar to that of $\alpha-$JR divergence. To further clarify, we demonstrate the separation performance of the $\alpha-$Rényi mutual divergence as ICA criterion, an experiment using the source mixed signal and an $\alpha = 1.6$ for a mutual divergence are shown in Fig.(7.8), as well as a comparison of 1.6−Rényi mutual divergence and mutual

information given in Fig.(7.9). From these two figures, we clearly see that $\alpha-$Rényi mutual divergence ($1 < \alpha < 2$) is a better criterion than mutual information, and holds an advantage over α-JR divergence whose efficient estimation presents some challenging issues.

7.5 Approximation of $\alpha-$Rényi Mutual Divergence

7.5.1 Introduction

In this Section, we investigate the estimation of Rényi mutual information using the relative frequencies calculated on cells of adaptive partitions of $\mathcal{R}_n$ of $X \times Y$. This is a generalization of a non-parametric estimation of mutual information proved by [86] and further implemented by [19].

As described in Section 7.3.1, Rényi divergence aims at measuring the distance between two probability density functions and is given by

$$RD_\alpha(p,q) = \frac{1}{\alpha - 1} \ln \left(\int p^\alpha(x) q^{1-\alpha}(x) dx \right), \tag{7.5.1}$$

with $0 < \alpha < 2$. It is identically 0 if the random variables are equal in distribution. The vanishing property of Rényi divergence is also equivalent, and as noted earlier, to independence of two random variables when their joint density and their marginal distributions are invoked. To proceed with the description of a non-parametric alternative method to estimate $\alpha-$Rényi mutual divergence of two random variables in this case, we state the following two theorems in the following.

7.5.2 Approximation Theorems of $\alpha-$Rényi mutual divergence

We consider a pair of random variables $\xi,\ \eta$ taking values in a measurable space $(X \times Y, S_X \times S_Y)$ with probability distributions $P_\xi(.)$ and $P_\eta(.)$ respectively. Assume the joint distribution $P_{\xi,\eta}(.)$ of $\xi,\ \eta$ is absolutely continuous with respect to the product distribution $P_\xi \times P_\eta(.)$, then from Radon-Nikodym theorem, there exists a function $a_{\xi,\eta}(x,y)$, assuming finite nonnegative values and measurable relative to the σ-algebra $S_X \times S_Y$, such that for all $B \in S_X \times S_Y$, the probability $P_{\xi,\eta}(B)$ is given by the integral of $a_{\xi,\eta}(x,y)$ over B with

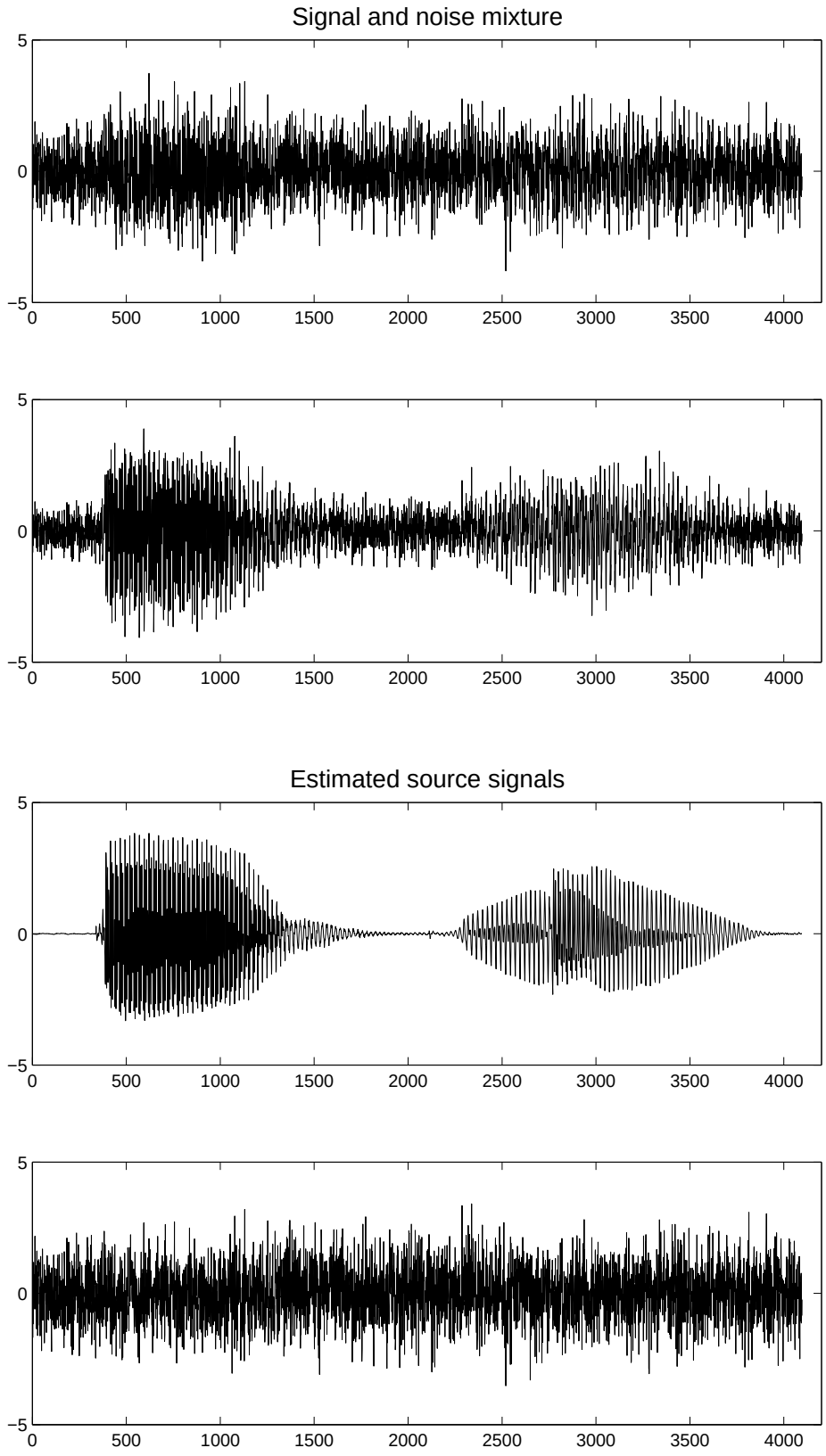

Figure 7.8: Mixed signals and its separation using 1.6−Rényi mutual divergence.

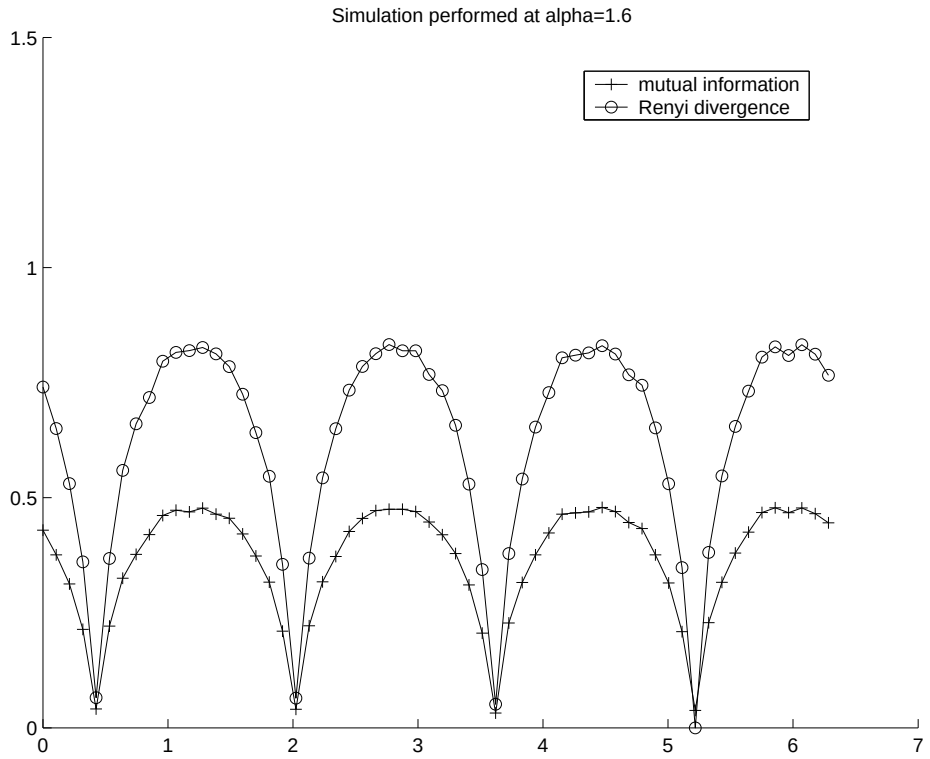

Figure 7.9: 1.6−Rényi mutual divergence measure compared to mutual information

respect to the measure $P_\xi \times P_\eta(.)$:

$$P_{\xi,\eta}(B) = \int_B a_{\xi,\eta}(x,y) P_\xi \times P_\eta(dx,dy) \tag{7.5.2}$$

The quantity $a_{\xi,\eta}(x,y)$ is called the density of the measure $P_{\xi\eta}$ with respect to the measure $P_\xi \times P_\eta(.)$ and is conventionally denoted by

$$a_{\xi,\eta}(x,y) = \frac{dP_{\xi\eta}(.)}{dP_\xi \times P_\eta(.)}. \tag{7.5.3}$$

If we assume probabilities $P_{\xi\eta}, P_\xi, P_\eta$ have probability densities, namely, $dP_{\xi\eta}(.) = h(x,y)dxdy$, $dP_\xi(x) = f(x)dx$, $dP_\eta(y) = g(y)dy$, from the definition of $\alpha-$Rényi mutual divergence of random variables X and Y , we have

$$\begin{aligned} MD_\alpha(X,Y) &= \frac{1}{\alpha-1} \ln \int_{X\times Y} h^\alpha(x,y) \left(f(x)g(y)\right)^{1-\alpha} dxdy \\ &= \frac{1}{\alpha-1} \ln \int_{X\times Y} a_{\xi,\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy). \end{aligned} \tag{7.5.4}$$

We consider the approximation of $\alpha-$Rényi mutual divergence based on a sequence of finite partition $\mathcal{C}^{(k)} = \{C_i^{(k)}\}$ of $X \times Y$ with the property that $\mathcal{C}^{(k_1)}$ is a finer partition than $\mathcal{C}^{(k_2)}$ when $k_1 > k_2$. A finite partition $\mathcal{C}$ of $X \times Y$ is $\mathcal{C} = \{C_i\}_{i=1,\cdots,n}$ such that $\bigcup_{i=1}^n C_i = X \times Y$, where $C_i,\ i = 1,\cdots,n$ are subsets of $X \times Y$ such that $C_i \bigcap C_j = \varnothing,\ i \neq j$. Each set C_i is called a cell of the finite partition $\mathcal{C}$. A partition $\mathcal{C}^{(1)}$ is a refinement (finer or nest partition) of another partition $\mathcal{C}^{(2)}$, if for each $C_i^{(1)} \in \mathcal{C}^{(1)}$, there exists a $C_j^{(2)} \in \mathcal{C}^{(2)}$ such that $C_i^{(1)} \subset C_j^{(2)}$. $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}$ are called nested partitions. An example of nested partitions is shown in Fig. (7.10).

Theorem 7.5.1. *For any partition $\mathcal{C} = (C_1,\cdots,C_n)$ of $X \times Y$, we have the following conclusions for different cases of α,*

Case. 1. $0 < \alpha < 1$,

$$\int_{X\times Y} a_{\xi,\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) = \inf_{\mathcal{C}} \sum_i P_{\xi\eta}^\alpha(C_i) (P_\xi \times P_\eta(C_i))^{1-\alpha} \tag{7.5.5}$$

Case. 2. $1 < \alpha \leq 2$,

$$\int_{X\times Y} a_{\xi,\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) = \sup_{\mathcal{C}} \sum_i P_{\xi\eta}^\alpha(C_i) (P_\xi \times P_\eta(C_i))^{1-\alpha} \tag{7.5.6}$$

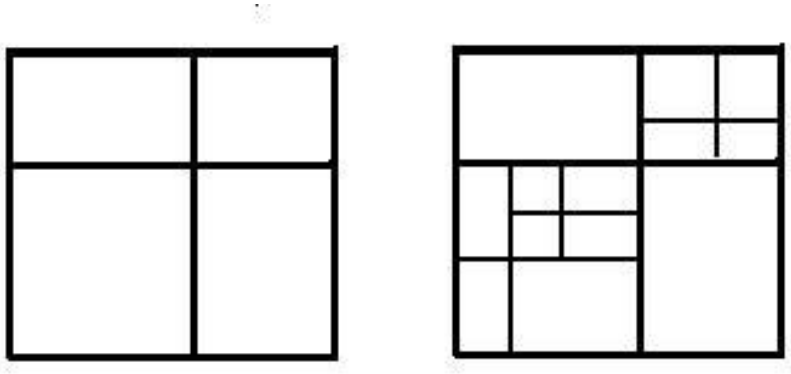

Figure 7.10: Example: Two nested partitions of $\mathcal{R}^2$

To practically implement Theorem (7.5.1) , we need the following theorem such that we can use the nested cell technique to develop a algorithm to efficiently approximate $\alpha-$Rényi mutual divergence by considering dependent data.

Theorem 7.5.2. *For a set of nested partitions $\mathcal{C}^{(k)}$ of $X \times Y$, we have*

Case 1: for $0 < \alpha < 1$,

$$\inf_{\mathcal{C}} \sum_i P_{\xi\eta}^{\alpha}(C_i)(P_\xi \times P_\eta(C_i))^{1-\alpha} = \lim_{k\to\infty} \sum_{C_i = A_i \times B_j \in \mathcal{C}^{(k)}} P_{\xi\eta}^{\alpha}(A_i \times B_j)(P_\xi(A_i) \times P_\eta(B_j))^{1-\alpha} \quad (7.5.7)$$

Case 2: for $1 < \alpha \leq 2$,

$$\sup_{\mathcal{C}} \sum_i P_{\xi\eta}^{\alpha}(C_i)(P_\xi \times P_\eta(C_i))^{1-\alpha} = \lim_{k\to\infty} \sum_{C_i = A_i \times B_j \in \mathcal{C}^{(k)}} P_{\xi\eta}^{\alpha}(A_i \times B_j)(P_\xi(A_i) \times P_\eta(B_j))^{1-\alpha} \quad (7.5.8)$$

The proof details for Theorems (7.5.1) and (7.5.2) are deferred to Appends A. and B.

Theorems (7.5.1) and (7.5.2) provide us with an approach to estimate α-Rényi divergence by appropriately dissecting dependent data space into sufficiently small cells that in the limit achieve the "inf" of the product given in Eq. (7.5.8).

To illustrate such a procedure, we generate a sequence of samples from a pair of correlated Bi-normally distributed random variables (X, Y), with the following joint probability density function

$$f(x, y) = \frac{1}{[2\pi Det(\Sigma)]^{1/2}} exp\left\{-\boldsymbol{v}\boldsymbol{\Sigma}^{-1}\boldsymbol{v}^{\tau}/2\right\}$$

where $\boldsymbol{v} = (x - \mu_x, y - \mu_y)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & r\sigma_x\sigma_y \\ r\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix},$$

and where r is the correlation coefficient of X and Y, and with marginal normal densities $f(x), f(y)$ with mean values μ_x, μ_y and variances σ_x^2, σ_y^2.

To this case, we may also compute the theoretical Rényi divergence and show it to be

$$\begin{aligned} & R^{\alpha}(X,Y) \\ = \; & \frac{1}{\alpha-1}\log\left(\int f(x,y)^{\alpha}(f(x)f(y))^{1-\alpha}dxdy\right) \\ = \; & \frac{2-\alpha}{2\alpha-2}\log(1-r^2) \\ & -\frac{1}{2\alpha-2}\log((1-(1-\alpha)r^2)^2-\alpha^2r^2). \end{aligned} \tag{7.5.9}$$

A comparison of our estimate and of the theoretical value are shown in Fig. (7.11), and demonstrates that an arbitrarily accurate estimate is achievable.

7.6 Conclusion

We have proposed two more robust information measure demonstrated for ICA, $\alpha-$JR divergence and $\alpha-$Rényi mutual divergence, but as we have argued, $\alpha-$JR divergence has a potentially broader scope of applicability depending on one's ability to appropriately interpret the probability measures and the related parameters in the JR divergence, this problem can be overcome by using Rényi divergence as we address the numerical complexity of this measure and propose a non-parametric alternative implementation.

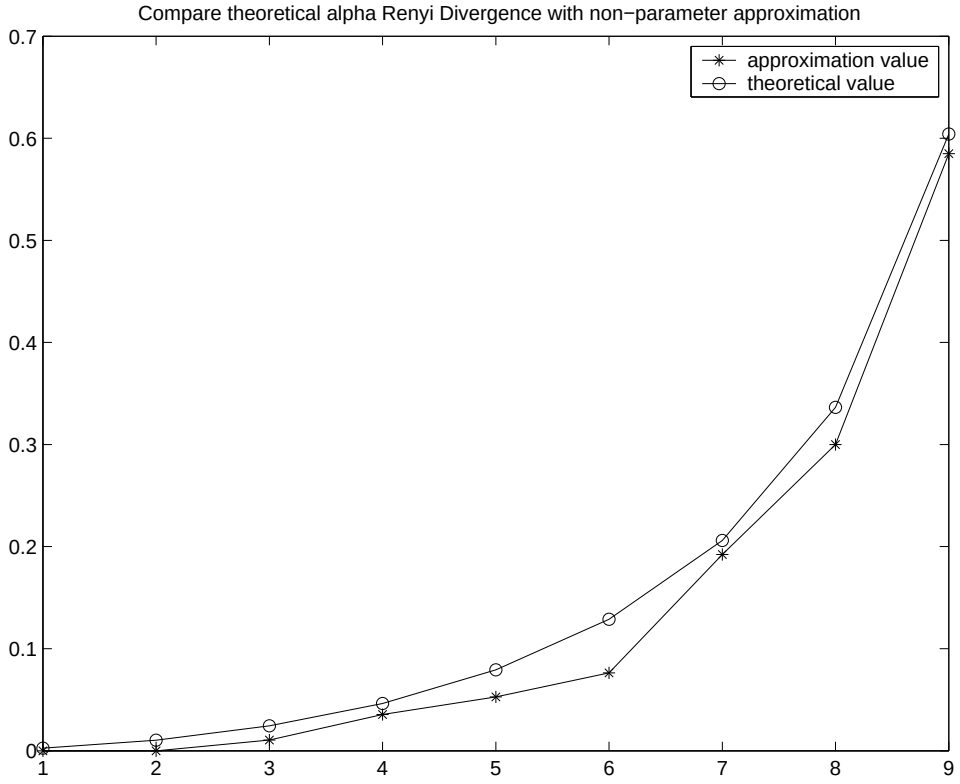

Figure 7.11: Approximated 0.5−Rényi mutual divergence and its exact theoretical value

7.7 Appendix A.

We only prove case 1 of Theorem (7.5.1), which is an immediate result of the following Lemma (7.7.1) and Lemma (7.7.2). Case 2 may be similarly proved.

If probabilities $P_{\xi\eta}(x, y)$ and $P_\xi \times P_\eta$ have corresponding probability densities $p_{\xi\eta}(x, y)$ and $p_\xi(x)p_\eta(y)$, we see that

$$a_{\xi,\eta}(x, y) = \frac{p_{\xi\eta}(x, y)}{p_\xi(x)p_\eta(y)} \tag{7.7.1}$$

and therefore Rényi mutual divergence may be obtained as

$$\int_{X\times Y} p^\alpha_{\xi,\eta}(x, y)(p_\xi(x)p_\eta(y))^{1-\alpha}dxdy = \frac{1}{\alpha - 1}\ln\int_{X\times Y} a_{\xi,\eta}(x, y)^{\alpha-1}P_{\xi,\eta}(dx, dy) \tag{7.7.2}$$

we have the following two inequalities regarding the approximation of the integral in the above equation, namely,

Lemma 7.7.1. *For $0 < \alpha < 1$, and any partitions of $X \times Y$,*

$$\int_{X\times Y} a_{\xi,\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) \leq \inf_{\mathcal{C}} \sum_i P_{\xi\eta}^{\alpha}(C_i)(P_\xi \times P_\eta(C_i))^{1-\alpha} \tag{7.7.3}$$

Proof: Taking a function $g(x) = x^\alpha$, which is concave when $0 < \alpha < 1$ (see Fig.(7.12), hence $-g(x)$ be a convex function), by Jensen's inequality, we have, for every probability distribution $F(x)$,

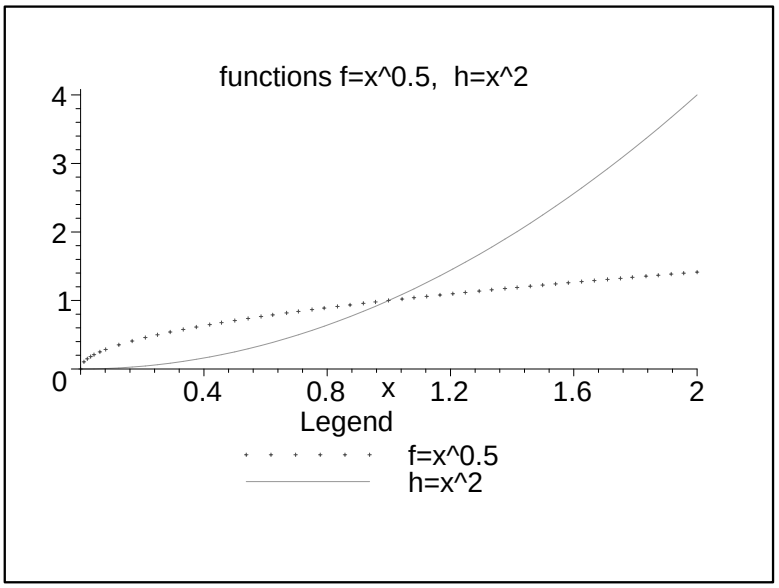

Figure 7.12: Functions of $f = x^\alpha$, with $0 < \alpha < 1$ and $1 < \alpha \leq 2$

$$\int_0^\infty x^\alpha dF(x) \leq \left(\int_0^\infty x dF(x) \right)^\alpha. \tag{7.7.4}$$

Now fixing a set $B \in S_X \times S_Y$ for which $P_\xi \times P_\eta(B) > 0$, and define a probability distribution of an event $A_u = (a_{\xi\eta}(x,y) < u)$ constrained to B

$$\begin{aligned} F_B(u) &= P_\xi \times P_\eta\{A_u|B\} \\ &= \frac{P_\xi \times P_\eta\{(a_{\xi\eta}(x,y) < u) \bigcap B\}}{P_\xi \times P_\eta(B)}, \end{aligned} \tag{7.7.5}$$

we have

$$\int_0^\infty u dF_B(u) = \frac{1}{P_\xi \times P_\eta(B)} \int_B a_{\xi\eta}(x,y) P_\xi \times P_\eta(dx,dy) = \frac{P_{\xi\eta}(B)}{P_\xi \times P_\eta(B)} \tag{7.7.6}$$

and

$$\int_0^\infty u^\alpha dF_B(u) = \frac{1}{P_\xi \times P_\eta(B)} \int_B a_{\xi\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) \tag{7.7.7}$$

Applying Eq. (7.7.4) to Eq. (7.7.7) shows that, for $\forall\ B \in S_X \times S_Y$ with $P_\xi \times P_\eta(B) > 0$, we have

$$\begin{aligned} & \int_B a_{\xi\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) \\ = & \frac{1}{P_\xi \times P_\eta(B)} \int_0^\infty u^\alpha dF_B(u) \\ \leq & P_\xi \times P_\eta(B) \left(\frac{P_{\xi\eta}(B)}{P_\xi \times P_\eta(B)} \right)^\alpha \\ = & (P_{\xi\eta}^\alpha(B))(P_\xi \times P_\eta(B))^{1-\alpha}. \end{aligned} \tag{7.7.8}$$

We can see that Eq. (7.7.8) is also true when $P_\xi \times P_\eta(B) = 0$. Now we consider a certain dissection $\{C_i\}$ of the space $X \times Y$. From Eq. (7.7.8), we have

$$\begin{aligned} I(C_1, \cdots, C_n) & \triangleq \sum_{i=1}^n (P_{\xi\eta}^\alpha(C_i))(P_\xi \times P_\eta(C_i))^{1-\alpha} \\ & \geq \sum_{i=1}^n \int_{C_i} a_{\xi\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) \\ & = \int_{X\times Y} a_{\xi\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) \end{aligned} \tag{7.7.9}$$

since dissection $\{C_i\}$ of the space $X \times Y$ is arbitrarily chosen, we see that Eq. (7.7.3) is established.

Lemma 7.7.2. *For $0 < \alpha < 1$, and any partitions of $X \times Y$,*

$$\int_{X\times Y} a_{\xi,\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) \geq \inf_{\mathcal{C}} \sum_{i,j} P_{\xi\eta}^\alpha(C_i)(P_\xi \times P_\eta(C_i))^{1-\alpha} \tag{7.7.10}$$

Proof: For $\forall\ \epsilon > 0$, since $x^\alpha \to 0$ as $x \to 0$ and $x^{\alpha-1} \to 0$ as $x \to \infty$ for $0 < \alpha < 1$, we can choose a constant K small enough so that,

$$0 \leq P_{\xi\eta}^\alpha \{a_{\xi\eta}^{\alpha-1}(x,y) \leq K\} \leq \frac{\epsilon}{2}. \tag{7.7.11}$$

Now consider a set $\{a_{\xi\eta}^{\alpha-1}(x,y) > K\}$, which can be represented in the form of a sum of nonintersecting sets $C_i \in S_X \times S_Y, i = 1, \cdots, n$ so that for all i

$$\underline{h}_i = \inf_{(x,y)\in C_i} a_{\xi\eta}(x,y), \quad \overline{h}_i = \sup_{(x,y)\in C_i} a_{\xi\eta}(x,y)$$

we have

$$(\underline{h}_i)^{\alpha-1} - (\overline{h}_i)^{\alpha-1} \leq \frac{\epsilon}{2} \tag{7.7.12}$$

From Eq. (7.7.6), let $B = C_i$, we have that, for $\forall\ i$

$$\underline{h}_i \leq \frac{P_{\xi\eta}(C_i)}{P_\xi \times P_\eta(C_i)} \leq \overline{h}_i, \tag{7.7.13}$$

and yielding

$$(\overline{h}_i)^{\alpha-1} \leq \left(\frac{P_{\xi\eta}(C_i)}{P_\xi \times P_\eta(C_i)}\right)^{\alpha-1} \leq (\underline{h}_i)^{\alpha-1}, \quad 0 < \alpha < 1. \tag{7.7.14}$$

Further, by the definition of $\underline{h}_i, \overline{h}_i$, we have

$$P_{\xi\eta}(C_i)(\overline{h}_i)^{\alpha-1} \leq \int_{C_i} a_{\xi\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) \leq P_{\xi\eta}(C_i)(\underline{h}_i)^{\alpha-1}. \tag{7.7.15}$$

From Eq. (7.7.14) and Eq. (7.7.15) we see that

$$\left| P_{\xi\eta}^{\alpha}(C_i)(P_\xi \times P_\eta(C_i))^{1-\alpha} - \int_{C_i} a_{\xi\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) \right| \leq [(\underline{h}_i)^{\alpha-1} - (\overline{h}_i)^{\alpha-1}] P_{\xi\eta}(C_i), \tag{7.7.16}$$

summing the inequalities Eq. (7.7.16) over i and using Eq. (7.7.12), we have

$$\left| I(C_1, \cdots, C_n) - \int_{\{a_{\xi\eta}^{\alpha-1} > K\}} a_{\xi\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) \right| \leq \frac{\epsilon}{2}. \tag{7.7.17}$$

Let $C_{n+1} = \{a_{\xi\eta}^{\alpha-1} \leq K\}$, we see that the system of sets $C_1, \cdots, C_n, C_{n+1}$ form a dissection of the space $X \times Y$. From Eq. (7.7.17) we have that

$$\begin{aligned} & I(C_1, \cdots, C_n, C_{n+1}) = I(C_1, \cdots, C_n) + (P_{\xi\eta}^{\alpha}(C_{n+1})(P_\xi \times P_\eta(C_{n+1}))^{1-\alpha}) \\ \leq & \int_{\{a_{\xi\eta}^{\alpha-1} > K\}} a_{\xi\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) + P_{\xi\eta}^{\alpha}(C_{n+1})(P_\xi \times P_\eta(C_{n+1}))^{1-\alpha} + \frac{\epsilon}{2} \end{aligned} \tag{7.7.18}$$

Since $C_{n+1} = \{a_{\xi\eta}^{\alpha-1} \leq K\}$, It is clear that

$$\begin{aligned} & P_{\xi\eta}^{\alpha}(C_{n+1})(P_\xi \times P_\eta(C_{n+1}))^{1-\alpha} \\ = & P_{\xi\eta}^{\alpha}(\{a_{\xi\eta}^{\alpha-1} \leq K\})(P_\xi \times P_\eta(\{a_{\xi\eta}^{\alpha-1} \leq K\}))^{(1-\alpha)} \\ \leq & P_{\xi\eta}^{\alpha}(\{a_{\xi\eta}^{\alpha-1} \leq K\}) < \frac{\epsilon}{2} \end{aligned} \tag{7.7.19}$$

Combining this inequality with inequality (7.7.11), we see that

$$\inf I(C_1, \cdots, C_{n+1}) \leq I(C_1, \cdots, C_{n+1}) \leq \int_{\{a_{\xi\eta}^{\alpha-1} > K\}} a_{\xi\eta}^{\alpha-1}(x,y) P_{\xi\eta}(dx,dy) + \epsilon, \tag{7.7.20}$$

which in light of the fact that ϵ is arbitrary and K may be chosen small enough, we can see that Eq. (7.7.10) is established.

7.8 Appendix B

Similarly, we only need to prove Case 1 of Theorem 7.5.2(as case 2 is similarly proved), where we need the Lemmas (3) and (4)

Lemma 3. *For* $0 \leq x_1,\ x_2,\ y_1,\ y_2 \leq 1$, *we have the following inequality for* $0 < \alpha < 1$,

$$x_1^\alpha y_1^{1-\alpha} + x_2^\alpha y_2^{1-\alpha} \leq (x_1 + x_2)^\alpha (y_1 + y_2)^{1-\alpha} \tag{7.8.1}$$

Proof: Let $f(x_1, x_2, y_1, y_2) = x_1^\alpha y_1^{1-\alpha} + x_2^\alpha y_2^{1-\alpha} - (x_1 + x2)^\alpha (y_1 + y_2)^{1-\alpha}$

Take partial derivative of $f(x_1, x_2, y_1, y_2)$ with respect to x_1, we will have

$$\begin{aligned} & f'_{x_1}(x_1, x_2, y_1, y_2) \\ = \ & \alpha x_1^{\alpha-1} y_1^{1-\alpha} - \alpha (x_1 + x_2)^{\alpha-1} (y_1 + y_2)^{1-\alpha} \\ = \ & \alpha \left[\left(\frac{x_1}{y_1} \right)^{\alpha-1} - \left(\frac{x_1 + x_2}{y_1 + y_2} \right)^{\alpha-1} \right] \end{aligned} \tag{7.8.2}$$

where, for $0 < \alpha < 1$, and $y_1,\ y_2 > 0$

$$f'_{x_1} = \begin{cases} < 0 & \text{if } \dfrac{x_1}{y_1} > \dfrac{x_1 + x_2}{y_1 + y_2}, \text{ namely } x_1 > \dfrac{y_1}{y_2} x_2 \\ \geq 0 & \text{if } \dfrac{x_1}{y_1} \leq \dfrac{x_1 + x_2}{y_1 + y_2}, \text{ namely } x_1 \leq \dfrac{y_1}{y_2} x_2 \end{cases}$$

This leads to the following inequalities for $\forall\ y_1,\ y_2 > 0$ when $1 \geq x_1 > \frac{y_1}{y_2} x_2$

$$f(1, x_2, y_1, y_2) \leq f(x_1, x_2, y_1, y_2) \leq f(\frac{y_1}{y_2} x_2, x_2, y_1, y_2),$$

and when $0 \leq x_1 \leq \frac{y_1}{y_2} x_2$

$$f(0, x_2, y_1, y_2) \leq f(x_1, x_2, y_1, y_2) \leq f(\frac{y_1}{y_2} x_2, x_2, y_1, y_2).$$

From these two inequalities, we have, for $0 \leq x_1,\ x_2 \leq 1, 0 < y_1,\ y_2 \leq 1$,

$$f(x_1, x_2, y_1, y_2) \leq f(\frac{y_1}{y_2} x_2, x_2, y_1, y_2)$$

and since

$$\begin{aligned} & f(\frac{y_1}{y_2}x_2, x_2, y_1, y_2) \\ = & \left(\frac{y_1}{y_2}x_2\right)^{\alpha} y_1^{1-\alpha} + x_2^{\alpha} y_2^{1-\alpha} - \left(\frac{y_1}{y_2}x_2 + x_2\right)^{\alpha} (y_1+y_2)^{1-\alpha} \\ = & \frac{x_2^{\alpha}(y_1+y_2)}{y_2^{\alpha}} - \frac{x_2^{\alpha}(y_1+y_2)^{\alpha}}{y_2^{\alpha}}(y_1+y_2)^{1-\alpha} \\ = & 0 \end{aligned} \tag{7.8.3}$$

(7.8.1) is immediately obtained for $y_1,\ y_2 > 0$, which is clearly true when $y_1 = 0$ or $y_2 = 0$.

With the help of Lemma (3), we have the following Lemma (4),

Lemma 4. *For any two nested partitions $\mathcal{C}^{(k)}, \mathcal{C}^{(l)}, k > l$ of $X \times Y$, we have for $0 < \alpha < 1$,*

$$\begin{aligned} & \sum_{C_i = A_i \times B_j \in \mathcal{C}^{(k)}} P_{\xi\eta}^{\alpha}(A_i \times B_j)(P_{\xi}(A_i) \times P_{\eta}(B_j))^{1-\alpha} \\ \leq & \sum_{C_i = A_i \times B_j \in \mathcal{C}^{(l)}} P_{\xi\eta}^{\alpha}(A_i \times B_j)(P_{\xi}(A_i) \times P_{\eta}(B_j))^{1-\alpha} \end{aligned} \tag{7.8.4}$$

Proof:

We know that for any cell $A \times B \in \mathcal{C}^{(l)}$, $A \times B$ can be written as

$$A \times B = \sum_{m=1}^{M} A_m \times B_m$$

To prove Lemma 4, we only need to show that

$$P_{\xi\eta}^{\alpha}(A \times B)(P_{\xi}(A) \times P_{\eta}(B))^{1-\alpha} \geq \sum_{m=1}^{M} P_{\xi\eta}^{\alpha}(A_m \times B_m)(P_{\xi}(A_m) \times P_{\eta}(B_m))^{1-\alpha}$$

If $M = 2$

$$\sum_{m=1}^{2} P_{\xi\eta}^{\alpha}(A_m \times B_m)(P_{\xi}(A_m) \times P_{\eta}(B_m))^{1-\alpha} \tag{7.8.5}$$

$$\leq [P_{\xi\eta}^{\alpha}(A_1 \times B_1) + P_{\xi\eta}^{\alpha}(A_2 \times B_2)][P_{\xi}(A_1) \times P_{\eta}(B_1) + P_{\xi}(A_2) \times P_{\eta}(B_2)]^{1-\alpha} \tag{7.8.6}$$

In the case that $A = A_1 + A_2, B = B_1 = B_2$, from Lemma (3), we have

$$
\begin{aligned}
& [P^{\alpha}_{\xi\eta}(A_1 \times B_1) + P^{\alpha}_{\xi\eta}(A_2 \times B_2)][P_{\xi}(A_1) \times P_{\eta}(B_1) + P_{\xi}(A_2) \times P_{\eta}(B_2)]^{1-\alpha} \\
\leq \quad & P^{\alpha}_{\xi\eta}(A \times B)[P_{\xi}(A) \times P_{\eta}(B)]^{1-\alpha} \qquad (7.8.7)
\end{aligned}
$$

This inequality can be repeat and proved that it is true for $A = A_1 + A_2, B = B_1 + B_2$, thus can be easily generated to a general number M, thus Lemma (4) follows.

Chapter 8

Possible Future Work

In this chapter, we overview briefly the contributions of this dissertation, we also present some possible further development to extend our work.

8.1 Summaries

Our first contribution in this work is the construction of a stochastic framework for nonlinear diffusions and its utilization to solve a long standing problem of an evolution stopping time. While the previously mention contribution provided a significant gain in denoising and in segmentation, it also need improves in texture preservation. In showing a direct connection between nonlinear diffusion and wavelet frame analysis filtering, efficient and texture preserving nonlinear techniques were developed.

While nonlinearities introducing in process were aimed at accounting for the various dependencies among the components of a signal/image, an alternative approach would be, and as developed in chapter 6 and 7, to use the underlying PDF's to find and separate independent components(ICA)

8.2 Possible Future Research

In this section, we list several potential topics related to this thesis that might constitute good leads for future research.

Among the many extensions one might pursue is an 8-neighbor transition scenario in the Markov chain. Another interesting variation on the theme is a two-step transition random walk which may also ultimately be driven to a continuous space setting.

In the nonlinear wavelet frame setting, wavelets, with a higher order of vanishing moments are expected to yield a better performance, while the analytical tractability of the problem as resolved in chapter 5 remains an open problem.

While theoretical not limited in the dimensionality of the ICA problem, our 1-D development is in real need to be extended to ultimately address two and higher dimensions for applications ranging from denoising to feature extraction to classification and recognition.

Bibliography

[1] B. Ma A. O. Hero and O. Michel, *Alpha-divergence for image indexing and retrieval*, Preprint (2000).

[2] S. I. Amari and A. Cichocki, *Adaptive blind signal processing - neural network approaches*, Proceedings of the IEEE **10** (1998), 2026–2048.

[3] L. Arnold, *Stochastic differential equations: Theory and applications*, John Wiley and Sons, New York, London, Sydney, Toronto, 1974.

[4] Y. Bao and H. Krim, *bridging scale-space to multiscale frame analysis*, ICASSP'01 Salt lake city.

[5] ———, *A new criterion to independent component analysis*, ICASSP'02, submitted.

[6] ———, *Upon bridging scale-space to multiscale frame analysis*, Wavelets in Signal and Image Analysis: From Theory to Practice, COMPUTATIONAL IMAGING AND VISION ,Volume 19, Chapter 6.

[7] A. J. Bell and T. J. Sejnowski, *An information-maximization approach to blind separation and blind deconvolution*, Neural computation **7(6)** (1995), 1004–1034.

[8] J. Canny, *A computational approach to edge detection*, IEEE Trans. on PAMI, vol. PAMI-8, No. 6 (Nov, 1986).

[9] J-F. Cardoso, *Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem*, In Proc. ICASSP'90, Albuquerque, NM, USA (1990), 2655–2658.

[10] ———, *Super-symmetric decomposition of the fourth-order cumulant tensor, blind identification of more sources than sensors*, In Proc. ICASSP'91 (1991), 3109–3112.

[11] J. F. Cardoso, *Iterative techniques for blind sources separation using only fourth order cumulants*, Eusipco, 1992, pp. 739–742.

[12] J-F. Cardoso, *Blind signal separation: statistical principles*, Proceedings of the IEEE,special issue on blind identification and estimation **90** (1998), 2009–2026.

[13] ———, *Infomax and maximum likelihood for source separation*, IEEE Letters on signal processing **4(4)** (Apr. 1997), 112–114.

[14] J-F. Cardoso and A. Souloumiac, *Blind beamforming for non gaussian signals*, IEEE Proceedings-F **140(6)** (Dec. 1993), 395–401.

[15] F. Catté, T. Coll, P. L. Lion, and J. M. Morel, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM K. Numer. Anal. , 29, 182-193 (1992).

[16] P. Comon, *Independent component analysis, a new concept?*, Signal Processing **36** (1994), no. 3, 287–314.

[17] ———, *Blind channel identification and extraction of more sources than sensors*, SPIE Conf. Adv. Sig. Proc. **VIII, San Diego** (22-24, July,1998), 2–13.

[18] P. Garrat D. T. Pham and C. Jutten, *Separation of a mixture of independent sources through a maximum likelihood approach*, In Proc. EUSIPCO (1992), 771–774.

[19] G. A. Darbellay and I. Vajda, *Estimation of the mutual information by an adaptive partitioning of the observation space*, IEEE Transactions on Information Theory **45** (1999), no. 5, 1315–1321.

[20] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Com. Pure and Appl. Math. **XLI** (1988), 909–996.

[21] I. Daubechies, *Ten lectures on wavelets*, SIAM, Philadephia, PA (1992).

[22] A. Hyv ärinen, *Fast and robust fixed-point algorithms for independent component analysis*, IEEE Trans. on Neural Networks **10** (May 1999), 626–634.

[23] D. L. Donoho and I. M Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, preprint Dept. Stat. , Stanford Univ., Jun. 1992.

[24] R. J. Duffin and A. C. Schaeffer, *A class of nonharmonic fourier series*, Trans. Amer. Math. Soc. **72** (1952), 341–366.

[25] R. Durrett, *Stochastic calculus: A practical introduction*, first ed., CRC Press, Florida, 1997.

[26] E. Dynkin and A. Yushkevich, *Controlled markov processes. series of comprehensive studies in mathematics*, Springer-Verlag, 1975.

[27] E. B. Dynkin and A. A. Yushkevich, *Controlled markov processes*, Springer-Verlag Berlin-Heidelberg NewYork, 1975.

[28] W. H. Fleming and H. M. Soner, *Controlled markov processes and viscosity solutions*, Springer-Verlag Berlin-Heidelberg NewYork, 1992.

[29] A. Friedman, *Differential equations and applications*, Academic press, New York, San Francisco,London, 1976.

[30] J. H. Friedman and J. W. Tukey, *A projection pursuit algorithm for exploratory data analysis*, IEEE Trans. of Computers **c-23(9)** (1974), 881–890.

[31] D. Geman and G. Reynolds, *Constrained restoration and the recovery of discontinuities*, IEEE Transactions on PAMI **PMAI-14** (1992), no. 3, 367–382.

[32] S. Geman and D. Geman, *Stochastic relaxation, gibbs distributions, and the baysian restoration of images*, IEEE Transactions on PAMI **PMAI-6** (1984), no. 6, 721 –741.

[33] J. Geronimo, D. Hardin, and P. R. Massupust, *Fractal functions and wavelet expansions based on several scaling functions*, J. of Approx. Theory **78** (1994), 373–401.

[34] I. I. Gihman and A. V. Skorohod, *Stochastic differential equations*, Springer-Verlag, Ner York, Heidelberg, Berlin, 1972.

[35] A. B. Hamza and H. Krim, *Towards a unified view of estimation: Variational vs. statistical*, submitted to IEEE Trans. on Signal Processing (2001).

[36] H. H. Harman, *Modern factor analysis*, University of Chicago Press, 2nd edition, 1967.

[37] Y. He, A. B. Hamza, and H. Krim, *An information divergence measure with its application to image registration*, submitted to IEEE Transactions on Signal Processing, 2001.

[38] A. O. Hero and O. Michel, *Robust entropy estimation strategies based on edge weighted random graphs*, in Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE), San Diego, CA (July, 1998).

[39] ———, *Estimation of rényi information divergence via pruned minimal spanning trees*, in IEEE Trans. on Inform. Theory **IT-45** (Sept, 1999), no. 6, 1921–1939.

[40] H. P. Hiriyannaiah, G. L. Bilbro, W. Snyder, and R. C. Mann, *Restoration of piecewise-constant images by mean-field annealing*, Journal of the optical society of America A, vol. 6 No. 12 (Dec. ,1989).

[41] P. J. Huber, *Projection pursuit*, Ann. of Statist. **13(2)** (1985), 435–475.

[42] A. Hyvärinen, *Survey on independent component analysis*, Helsinki University of Technology.

[43] ———, *New approximations of differential entropy for independent component analysis and projection pursuit*, In Advances in Neural Information Processing Systems 10 (1998), 273–279.

[44] L. Wang R. Vigario J. Karhunen, E. Oja and J. Joutsensalo, *A class of neural networks for independent component analysis*, IEEE Trans. on Neural Networks **8** (1997), no. 3, 486–504.

[45] S. Jaffard, *Pointwise smoothness, two-microlocalization and wavelet coefficients*, Publications Matemátiques **35** (1991), 155–168.

[46] I. T. Jolliffe, *Principle component analysis*, Springer-Verlag, 1986.

[47] C. Jutten and J. Herault, *Blind separation of sources, part i: An adaptive algorithm based on neuromimatic architecture*, Signal processing **24** (July 1991), 1–10.

[48] J. J. Koendrink, *The structure of images*, Biol. Cybern. **50** (1987), 363–370.

[49] H. Krim and Y. Bao, *Nonlinear diffusion: A probabilistic view*, ICIP, Kobe, Japan, 1999.

[50] ———, *Smart nonlinear diffusion: A probabilistic approach*, Journal paper. Submitted to PAMI (2002).

[51] H. Krim, S. Mallat, D. Donoho, and A. Willsky, *Best basis algorithm for signal enhancement*, ICASSP'95 (Detroit, MI), IEEE, May 1995.

[52] H. Krim and I. C. Schick, *Minimax description length for signal denoising and optimized representation*, IEEE Trans. on Information Theory (1999), IEEE Trans. on IT Special Issue, Eds. H. Krim, W. Willinger, A. Iouditski and D. Tse.

[53] S. Kullback and R. Liebler, *On information and sufficiency*, Ann. Math. Statist. **22** (1951), 79–86.

[54] H. Kunita, *Stochastic flows and stochastic differential equations*, Academic press, Cambridge, New York,Port Chester, Melbourne ,London, 1976.

[55] ———, *Stochastic differential equations and stochastic flows of diffeomorphisms*, Lecture Notes Math. **1097** (1978), 143–303.

[56] Kushner and A. A. Yushkevich, *Numerical controlled diffusion*, Springer-Verlag Berlin-Heidelberg NewYork, 1990.

[57] H. J. Kushner and P. G. Dupuis, *Numerical methods for stochastic control problems in continuous time*, first ed., Applications of Mathematics, Springer-Verlag, 1992.

[58] P. L. Lions L. Alvarez and J. M. Morel, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM J. Num. Anal. **29** (1992), no. 3.

[59] L. De Lathauwer, P. Comon, B. De Moor, and J. Vandewalle, *Ica algorithms for 3 sources and 2 sensors*, IEEE Sig. Proc. Workshop on Higher-Order Statistics **Caesarea, Israel** (14-16, June,1999), 116–120.

[60] Stan Z. Li, *Close-form solution and parameter selection for convex minimization-based edge-preserving smoothing*, IEEE Trans. on PAMI, vol. 20, No. 9 (Sept, 1998).

[61] J. Lin, *Divergence measures based on the shannon entropy*, IEEE Trans. on Information Theory **37** (1991), 145–151.

[62] D. G. Luenberger, *Optimization by vector space methods*, J. Wiley, New York, London, 1968.

[63] G. Sapiro M. J. Black and D. H. Marimont, *Robust anisotropic diffusion*, IEEE Transactions on Image Processing **7** (1998), no. 3, 421–432.

[64] S. Mallat, *Multiresolution approximations and wavelet orthonormal bases of* $l^2(\mathcal{R})$, Trans. Amer. Math. Soc. **315** (1989), 69–87.

[65] S. Mallat, *A theory for multiresolution signal decomposition: the wavelet representation*, IEEE Trans. Patt. Anal. Mach. Intell. **PAMI-11** (1989), 674–693.

[66] _____, *A wavelet tour of signal processing*, Academic Press, Boston, 1997.

[67] S. Mallat and W. L. Hwang, *Singularity detection and processing with wavelets*, IEEE Transactions on Information Theory **38(2)** (1992), 617–643.

[68] H. S. Malvar, *The lot: a link between block transform coding and multirate filter banks*, IEEE International Symposium on Circuits and Systems, Helsinki, Finland.

[69] Y. Meyer, *Wavelets and applications*, first ed., SIAM, Philadelphia, 1992.

[70] E. Oja, *The nonlinear pca leaning rule in independent component analysis*, Neurocomputing **17** (1997), no. 1, 25–46.

[71] B. Oksendal, *Stochastic differential equations: An introduction with applications*, Springer-Verlag, Berlin, 1992.

[72] A. Papoulis, *Probability, random variables and stochastic processes*, Mc-Graw-Hill, N. Y., 1984.

[73] P. Perona, *Orientation diffusions*, IEEE Transactions on Image Processing **7** (1998), no. 3, 457–467.

[74] P. Perona and J. Malik, *A network for multiscale image segmentation*, IEEE Int. Symp. on Circuits and Systems (Helsinki), June 1988, pp. 2565–2568.

[75] ———, *Scale-space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Analysis and Machine Intelligence **12** (1990).

[76] D. T. Pham and P. Garat, *Blind separation of mixture of independent sources a quasi-maximum likelihood approach*, IEEE Trans. of Signal Processing **45(7)** (1997), 1712–1725.

[77] I. Pollak, A. S. Willsky, and H. Krim, *Image segmentation and edge enhancement with stabilized inverse diffusion equations*, IEEE Trans. on Image Processing. Feb, 00.

[78] ———, *Image segmentation and edge enhancement with stabilized inverse diffusion equations*, IEEE Trans. on Image Processing (Feb, 2000).

[79] V. Wickerhauser R. R. Coifman, Y. Meyer, *Wavelet analysis and signal processing*, In Wavelets and their applications (Boston 1992), 153–178. Jones and Barlett Publishers.

[80] Y. Meyer R. R. Coifman, *Remarques sur l'analyse de fourier fentre*, C. R. Acad. Sci. Paris **312, srie I** (1991), 259–261.

[81] A. Rényi, *On measures of dependence*, Acta Math. Acad. Sc. Hungar. **10** (1959), 441–451.

[82] ———, *On measures of entropy and information*, Selected Papers of Alfred Rényi **2** (1976), 525–580.

[83] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D, 1992.

[84] N. Saito, *Local feature extraction and its applications using a library of bases*, Ph.D. thesis, Yale University, Dec. 1994.

[85] C. E. Shannon, *A mathematical theory of communication*, Bell Syst. Tech. J. 27, pt. I, pp. 379-423; pt. II, pp. 623-656 (1948).

[86] R. L. Doubrušhin, *Iterative techniques for blind source separation using only forth-order cumulants*, In Proc. EUSIPCO,Brussels, Belgium (1992), 739–742.

[87] W. Snyder, Y. Han, G. Bilbro, R. Whitaker, and S. Pizer, *Image relaxation: restoration and feature extraction*, IEEE Trans. on PAMI,Vol. 17,No. 6 (Jun. ,1995).

[88] Ph. Tchamitchian, *Biorthogonalité et théorie des opérateurs*, Revista Matemática Iberoamericana **3(2)** (1987), 163–189.

[89] B. M. ter Haar Romeney, *Geometry driven diffusion*, Dordecht: Kluwer Academic Publishers, 1994, Edited Book.

[90] S. A. Tewfik, *Simple regularity criteria for subdivision scheme*, SIAM J. Math. Anal. **23** (1992), 1544–1576.

[91] C. Torre and T. A. Poggio, *On edge detection*, IEEE Trans. on PAMI, vol. PAMI-8, No. 2 (Mar, 1986).

[92] P. P. Vaidyanathanm, *Multirate systems and filter banks*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[93] G. W. Wei, *Generalized perona-malik equation for image restoration*, IEEE Transactions on Image Processing **7** (1998), no. 3, 457–467.

[94] R. Whitaker and S. Pizer, *A multi-scale approach to nonuniform diffusion*, Image understanding, vol. 57 99-110 (Jan. ,1993).

[95] A. P. Witkin, *Scale space filtering*, Proc. Int. Joint Conf. on Artificial Intelligence (Karlsruhe, Germany), 1983, pp. 1019–1023.

[96] Yu-Li You, Wenyuan Xu, Allen Tannenbaum, and Mostafa Kaveh, *Behavioral analysis of anisotropic diffusion in image processing*, IEEE Trans. on Image Proc. Vol,5. No. 11 p1539-1552. 1996.

[97] A. L. Yuille and T. Poggio, *Scaling theorems for zero-crossings*, JOSA **A** (1985), no. 2.